Collège Doctoral du Languedoc-Roussillon

THÈSE

Pour obtenir le grade de

Docteur

Délivré par **Montpellier SupAgro**

Préparée au sein de l'école doctorale **SIBAGHE**
Et de l'unité de recherche **AGAP**

Spécialité : **Biologie intégrative des plantes**

Présentée par **Guillaume MARTIN**

Caractérisation des différences de structures chromosomiques dans l'espèce *Musa acuminata* par re-séquençage ; le cas de l'accession 'Pahang'

Soutenue le 18 décembre 2014 devant le jury composé de

M. Eric JENCZEWSKI, Chargé de recherche, INRA	Rapporteur
M. Dominique LAVENIER, Directeur de recherche, CNRS	Rapporteur
Mme Janice BRITTON-DAVIDIAN, Directeur de recherche, CNRS	Examinateur
Mme Angélique D'HONT, Chercheur – HDR, CIRAD	Directeur de thèse

Montpellier SupAgro

Doctorat réalisé au sein de l'équipe Structure et évolution des génomes
UMR Amélioration Génétique et Adaptation des Plantes méditerranéennes et tropicales
Campus Lavalette, Bât 3, TA A-108 / 03 - Avenue Agropolis - 34398 Montpellier

Remerciements

Tout d'abord, je remercie Eric Jenczewski et Dominique Lavenier, les rapporteurs de cette thèse, qui ont accepté de consacrer du temps à lire ce manuscrit. Je remercie également Janice Britton-Davidian qui fera partie du jury, assistera à ma soutenance et participera à la discussion qui s'en suivra.

Merci aux différentes personnes (Jean-François Dufayard, Jean-Christophe Glaszmann, Laurent Journot, Olivier Panaud, Manuel Ruiz, Patrice This) qui ont assisté à mes comités de thèse pour leurs conseils et commentaires.

Je tiens à remercier les membres de l'UMR AGAP et plus particulièrement les habitants du bâtiment 3 pour leur accueil, leurs bonjours, renseignements et discussions en tout genre.

Merci à Xavier Perrier pour ces discussions « Banane », commentaires et remarques. Merci à Jean-Pierre Jacquemoud pour ses conseils et discussions « optimisation de script ».

Je remercie également les membres de l'équipe ID pour leur aide et leurs conseils en bioinformatique ainsi que pour nous avoir permis d'utiliser leurs fauteuils et tables (selon la saison) pendant les pauses. Un merci tout particulier à Gaëtan pour avoir pris le temps de m'initier à Galaxy.

Merci à Angélique D'Hont de m'avoir fait confiance pour réaliser ce travail de thèse. Angélique, merci de m'avoir guidé et conseillé tout au long de ces 3 ans. Merci d'avoir passé tant de temps en corrections, commentaires, suggestions et remaniements du manuscrit. Comme tu le dis : « tout se finit, même mal ». J'aurai tendance à dire que cette fois ci, ça s'est plutôt bien terminé…

Plus largement, merci à tous les membres de l'équipe SEG pour leur accueil. Grâce à vous, je me suis senti membre à part entière de l'équipe. Merci d'avoir supporté mon humour décalé et d'avoir ri, ou du moins essayé, à mes blagues pas toujours très bonnes.

Merci à Cyril Jourda, mon colocataire de bureau, merci pour ces discussions scientifiques et autres. J'ai été ravi de partager mon bureau avec toi.

Merci à Olivier Garsmeur pour nos échanges et discussions « script », pêche, champignons et bulldog ou pitbull (je ne sais plus...).

Merci aussi à Catherine Hervouet et Céline Cardi pour leurs encouragements et discussions en tout genre.

Merci à Carine Charron pour ses commentaires, corrections et encouragements.

Merci à Françoise Carreel pour nos discussions distorsions de ségrégation, mort de gamètes et rouge bleu. Merci également pour tes commentaires, corrections et encouragements.

Merci à Nabila Yahiaoui, merci pour avoir pris un verre d'eau à ma place. Merci pour ces discussions de fin de journée, pour tes commentaires, corrections et encouragements.

Merci à Florence Chazot pour son aide dans le domaine administratif et ses encouragements.

Franc-Christophe, merci de ton encadrement au cours de cette thèse. Merci pour ces discussions science ou encore « r*épétitions abéliennes dans les mots sturmiens* ». Merci à toi et à Carole pour ces randos photos, fleurs, champignons, papillons. Merci d'avoir fait découvrir le Sud à un Breton. Carole, merci d'avoir pris de ton temps pour corriger mon orthographe plutôt défaillante.

Je tenais à remercier Abdelkader et Malika Aïnouche pour m'avoir permis de faire mes premiers pas dans la recherche.

Enfin, je tenais à remercier l'ensemble des membres de la classe politique, acteurs de la télé-réalité et autres personnes suffisamment médiatisées pour leur application presque religieuse à alimenter nos discussions du midi.

Table des matières

Liste des figures :

Liste des tableaux :

Résumé

Les cultivars de bananiers sont issus d'hybridations entre sous-espèces de *Musa acuminata* (génome A, 2n=2x=22, 1C=500-600 Mbp) et pour certains avec l'espèce *M. balbisiana* (génome B, 2n=2x=22, 1C=550 bp). Ces hybrides présentent une fertilité réduite, des méioses perturbées et de fortes distorsions de ségrégation de leurs chromosomes. Chez les bananiers, ces caractéristiques sont attribuées à des réarrangements chromosomiques entre espèces et sous espèces. Bien que recherchée pour la production de fruits sans graines, la fertilité très réduite des cultivars associée aux fortes distorsions de ségrégation, compliquent les analyses génétiques et les programmes d'amélioration variétale. Au cours de cette thèse, nous avons mis en place et testé de nouvelles approches, basées sur la récente disponibilité d'une séquence de référence du bananier et des technologies de séquençage haut-débit, pour caractériser ces différences de structures chromosomiques et comprendre leur impact sur les ségrégations chromosomiques. Ces approches ont nécessité l'amélioration de la séquence de référence du bananier. Pour cela, des outils bioinformatiques ont été développés. Ils sont applicables à d'autres génomes et modulables en fonction des données disponibles. Le nombre de scaffolds a été divisé par 5 et 90% de la séquence est maintenant ancré aux chromosomes. Les scaffolds correspondant au génome mitochondrial ont été identifiés et le génome chloroplastique a été assemblé et annoté. Des données de re-séquençage de l'accession 'Pahang', le parent de l'haploïde doublé utilisé pour produire la séquence de référence, et des données de génotypage dense de sa descendance ont été utilisées pour explorer l'origine des distorsions de ségrégation impactant les chromosomes 1 et 4. A l'aide de simulations, nous montrons qu'un modèle génique de sélection gamétique pourrait, seul, expliquer les ségrégations observées. D'autre part, les modèles impliquant une variation structurale ne peuvent expliquer les données observées qu'avec l'ajout d'allèles sous sélection. La couverture obtenue avec les données de re-séquençage, nous permet d'éliminer l'hypothèse de la duplication. L'analyse réalisée avec le pipeline de recherche de variations structurales que nous avons développé, ne permet pas, seule, de conclure quant à la présence d'une translocation. Toutefois, les profils de distorsion et de recombinaison, les figures d'appariements à la méiose et les données de re-séquençage nous orientent vers l'hypothèse d'une translocation réciproque en orientation inversée. Le test de notre pipeline pour comparer les génomes A et B du bananier, dont les différences de structure sont connues, montre que notre pipeline détecte directement les signatures de certaines variations

structurales mais que, pour d'autres, il ne détecte que des signatures partielles. Ces dernières peuvent néanmoins être très informatives en complément d'autres types d'informations comme la cartographie génétique et les analyses cytogénétiques. L'amélioration de la séquence de référence du génome A réalisée lors de cette thèse aura un impact important sur la qualité des analyses réalisées par la communauté. Les outils développés pourront être utilisés pour l'analyse d'autres accessions de bananier ou d'autres espèces.

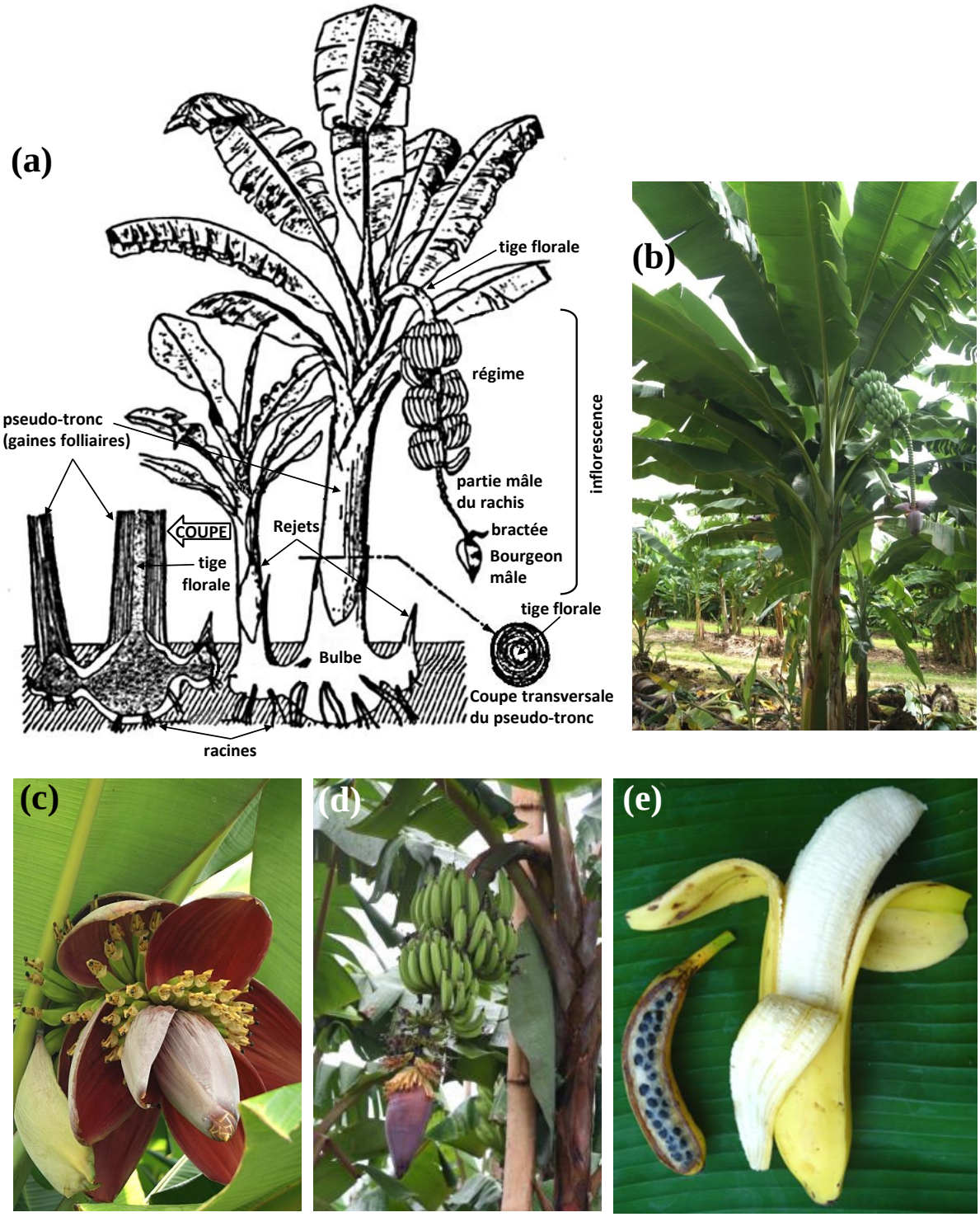


Figure 1: Le bananier. (a) Représentation schématique d'un bananier d'après (Champion, 1963). (b) Photographie d'un plan de bananier (G. Martin). (c) Jeune inflorescence de bananier découvrant deux rangées de fleurs à l'aisselle des bractées (G. Martin). (d) Régime de bananes avec du haut en bas, les mains femelles (bananes), les mains hermaphrodites, les mains mâles encore protégées par les bractées (F.C. Baurens, CIRAD). (e) Fruits de bananier séminifère (gauche) et parthénocarpique (droite) (A. D'Hont, CIRAD).

Description botanique

Les bananiers (Figure 1) sont des herbes géantes faisant généralement entre 2 et 5 mètres. La tige est souterraine (rhizome ou bulbe) et constitue la partie pérenne du bananier. C'est à partir de ce rhizome/bulbe que se développent les racines, les feuilles et l'inflorescence (Lassoudière, 2007). Avant la floraison, la partie aérienne du bananier n'est constituée que de feuilles dont les gaines foliaires sont imbriquées les unes dans les autres pour former le pseudo-tronc qui confère au bananier son aspect d'arbre. Les feuilles sont émises par le méristème terminal souterrain, au centre du pseudo-tronc. Les jeunes feuilles apparaissent au sommet du bouquet foliaire avec leur limbe enroulé autour de la nervure principale centrale formant une structure que l'on appelle le cigare. Pendant la floraison, le méristème terminal emprunte le même chemin que les feuilles à travers le pseudo-tronc en émettant une tige aérienne vraie appelée rachis ou hampe florale. Le bourgeon floral émerge au sommet de la couronne foliaire. Il est formé par l'imbrication de bractées caduques. A l'aisselle de chaque bractée, on trouve un groupe de fleurs organisé en deux rangées appelées mains. Chaque main comprend en moyenne une vingtaine de fleurs. Les premières mains qui apparaissent sont constituées de fleurs femelles ou quelques fois de fleurs hermaphrodites. Dans certains cas arrivent ensuite quelques mains de fleurs neutres (ni mâle ni femelle) (Simmonds, 1959). Les mains suivantes sont constituées de fleurs mâles. Ce sont les mains femelles qui donnent le régime de bananes. Leur nombre varie selon les espèces et variétés et les conditions environnementales. Le nombre de mains mâles varie en fonction des variétés et peut être indéfini.

A maturation du régime de fruits, la partie aérienne dégénère mais des rejets provenant de bourgeons secondaires, émis à l'aisselle des feuilles, permettent le maintien et la multiplication végétative de l'individu.

Le fruit du bananier est constitué de deux tissus : la peau, issue de la fusion de l'hypanthe avec l'exocarpe, qui constitue la couche de protection externe du fruit et la pulpe issue de l'endocarpe.

Parmi les bananiers, on identifie deux types : les bananiers séminifères (dits sauvages) et les bananiers parthénocarpiques (dits cultivés). Chez les bananiers séminifères, l'ovaire des fleurs femelles ne se développe en fruit que s'il y a eu fécondation et formation d'un zygote ; le fruit ainsi formé présente peu de pulpe et beaucoup de graines. Par contre, chez les bananiers parthénocarpiques, l'ovaire peut se développer en fruit sans graine avec une pulpe abondante dès la formation de la fleur femelle. Les graines de bananiers mesurent environ 5 mm de diamètre et sont en général extrêmement dures. Ces graines remplissent le fruit chez les bananiers séminifères. Chez les bananiers parthénocarpiques, les fruits sont généralement sans graine sauf après les fécondations forcées faites dans le cadre des programmes de création variétale (cf Introduction 1.3) où, suivant la fertilité résiduelle des cultivars, elles sont en plus ou moins grand nombre.

Le syndrome majeur de la domestication du bananier a été la perte de graines et le développement de la pulpe en l'absence de graines (parthénocarpie) (Simmonds, 1962). Cependant, les bananiers sont également exploités pour d'autres usages incluant la consommation des inflorescences en salade, l'utilisation des feuilles comme emballage alimentaire, l'utilisation des feuilles comme fourrage pour les animaux, l'utilisation des différentes parties du bananier en médecine, ou encore comme matériau de construction (Kennedy, 2009). Ces usages impliquent que des formes séminifères ont pu avoir été sélectionnées. La distinction entre bananiers séminifères et parthénocarpiques ne préjuge donc pas de leur niveau de domestication.

Systématique

Les bananiers sont des monocotylédones appartenant à l'ordre des Zingibérales et à la famille des *Musaceae*. Au sein des *Musaceae* trois genres sont identifiés : *Musella* (n=9), *Ensete* (n=9) et *Musa* (n=7/10/11) (Lassoudière, 2007; Li et al., 2010). Les bananiers du genre *Musella*, originaire du sud-ouest de la Chine, ne sont pas domestiqués et sont simplement cultivés à des fins ornementales ou religieuses en Asie (Pungetti et MacIvor, 2007). Les Ensete comprennent quelques espèces très proches et sont présents dans une vaste zone depuis l'Afrique jusqu'en Asie du sud-est. Ils sont cultivés pour leurs rhizomes riches en amidon en particulier en Éthiopie (Negash, 2001). Ils sont aussi utilisés comme plante ornementale en raison de leur robustesse. Les bananiers cultivés pour leurs fruits impliquent le genre *Musa*. Ce genre est divisé en 4 sections : *Autralimusa* (n=10), *Callimusa* (n=10), *Eumusa* (n=11) et *Rhodochlamys* (n=11) (Simmonds, 1959). La monophylie de ces différents genres et sections est remise en cause par des analyses phylogénétiques récentes qui suggèrent de constituer une section à 10 chromosomes et une à 11 (Christelova et al., 2011; Li et al., 2010; Wong et al., 2002). Au sein de la section *Australimusa*, distribuée en Asie du Sud-Est et en Océanie proche, seuls les 'Fehi' sont cultivés en Papouasie-Nouvelle Guinée et dans les îles du Pacifique. Ce sont des plants caractéristiques à régimes érigés et fruits riches en amidon et en carotène. Il n'existe pas de bananiers cultivés pour leurs fruits dans les sections *Callimusa* et *Rhodochlamys* qui ont plus un caractère ornemental. Les bananiers de la section *Eumusa* sont, si on inclut les formes cultivées, les plus répandus géographiquement (Simmonds, 1959). Cette section est constituée d'une vingtaine d'espèces à nombre chromosomique de base de 11. Deux de ces espèces, *Musa acuminata* (génome noté A, 1C=500-600 Mb) et *Musa balbisiana* (génome noté B, 1C=550Mb), sont à l'origine de la grande majorité des bananiers cultivés (Simmonds et Shepherd, 1955). L'espèce *Musa balbisiana* présente une aire de répartition assez étendue puisqu'on la retrouve dans toute la partie sud du continent asiatique (du nord de l'Inde jusqu'au sud de la Chine et peut-être aux Phillipines) (Cheesman, 1947) (**Figure 2**). On ne connaît pour cette espèce, par contraste avec *Musa acuminata*, qu'une faible diversité de formes (Cheesman, 1947) et ses représentants sont génétiquement très peu divergents (Carreel, 1994). L'espèce *Musa acuminata* a une vaste aire de répartition, puisqu'elle est présente dans toute l'Asie du Sud-Est continentale et insulaire. *Musa acuminata* a été classée en 6 à 9 sous-espèces inter-fertiles selon les auteurs (Carreel, 1994). La classification utilisée par Perrier et al. 2011 comprend 8 sous-espèces dont les aires de répartition sont bien identifiées (**Figure 2**):

M. a. banksii (Mueller)

M. a. burmannica (Simmonds)

M. a. errans (Blanco)

M. a. malaccensis (Ridley)

M. a. microcarpa (Beccari)

M. a. siamea (Simmonds)

M. a. truncata (Ridley)

M. a. zebrina (van Houtte).

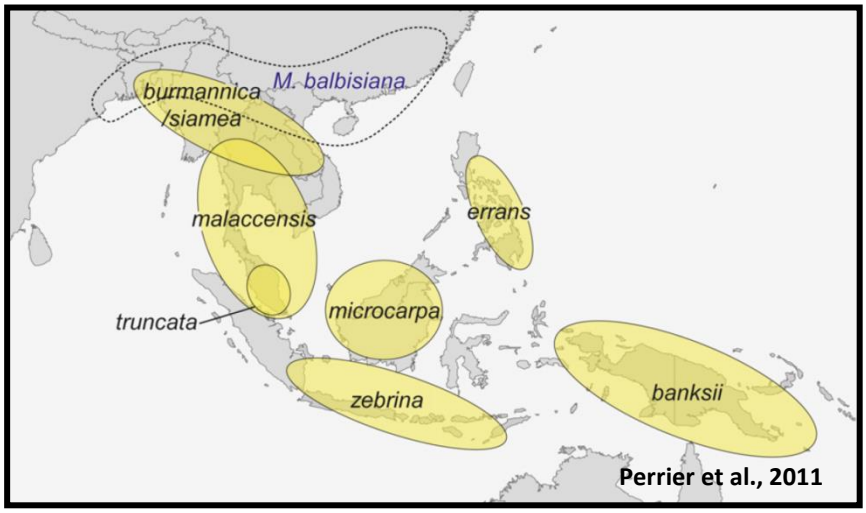


Figure 2 : Distribution géographique de l'espèce *Musa balbisiana* et des sous-espèces de *Musa acuminata*.

La dernière sous-espèce, *M. a. burmannicoïdes* (De Langhe), a été regroupée avec *M. a. burmannica*. La distinction entre sous-espèces répond de variations morphologiques majeures et a été confirmée par marqueurs moléculaires (Carreel, 1994; Hippolyte et al., 2012). Elle s'interprète par un isolement géographique dans ce milieu fortement insulaire et pourrait aboutir à terme à une vraie spéciation. La sous-espèce *M. a. banksii*, en particulier, présente essentiellement en Papouasie-Nouvelle-Guinée, a ainsi été proposée comme une espèce différente.

Introduction

1 - Origine et diversité des bananiers cultivés (parthénocarpiques)

1.1 Diversité et importance économique des bananiers cultivés

Les bananiers cultivés sont dérivés d'hybridations naturelles entre espèces et sous-espèces sauvages (cf encadrés 1 et 2) de *Musa*.

M. acuminata (génome A) est impliquée dans tous les cultivars et *M. balbisiana* (génome B) dans une partie d'entre eux (D'Hont et al., 2000; Simmonds, 1959; Simmonds et Shepherd, 1955). Ils ont été classés selon leur niveau de ploïdie et leur ressemblance morphologique avec les bananiers diploïdes sauvages (Cheesman, 1947; D'Hont et al., 2000; Simmonds et Shepherd, 1955). Ils peuvent être diploïdes (AA ou AB), cultivés principalement dans les zones d'origine. Le plus souvent, ils sont triploïdes (AAA, AAB, ABB) et présents dans toutes les zones de production. Ils sont rarement tétraploïdes *via* quelques produits des programmes d'amélioration.

Les bananiers cultivés jouent un rôle économique important dans les pays qui en exportent les fruits et/ou pour lesquels c'est un aliment de base. Les variétés les plus cultivées appartiennent au groupe des triploïdes AAA et sous-groupe Cavendish qui produit les bananes de type dessert, ce sont celles que nous mangeons en Europe. Ce sous-groupe comprend plus d'une vingtaine de cultivars (Lescot, 2014) divergents par quelques mutations somatiques (Raboin et al., 2005) et représente à lui seul plus de 50% de la production mondiale (61 millions de tonnes) de bananes et près de 97% des 19.6 millions de tonnes de bananes exportées. Les bananes dites Plantain (groupe AAB), constituent un sous-groupe de plusieurs dizaines de cultivars génétiquement très homogènes probablement mutants somatiques les uns des autres (Noyer et al., 2005). Ce sont des bananes à cuire qui représentent environ 15% de la production mondiale de bananes (Lescot, 2014). Ce sous-groupe est particulièrement cultivé en Afrique de l'Ouest et du Centre où il s'est diversifié pendant plus de 2000 ans et où il est un aliment de base. Un troisième grand sous-groupe de cultivars : les Lujugira-Mutika sont des triploïdes AAA cultivés en Afrique de l'Est et souvent appelés « Highland bananas ». Ils constituent une culture vivrière essentielle pour la sécurité alimentaire de ces pays (en particulier l'Ouganda, le Rwanda et le Burundi) où ils sont consommés crus ou cuits et servent même à faire de la bière.

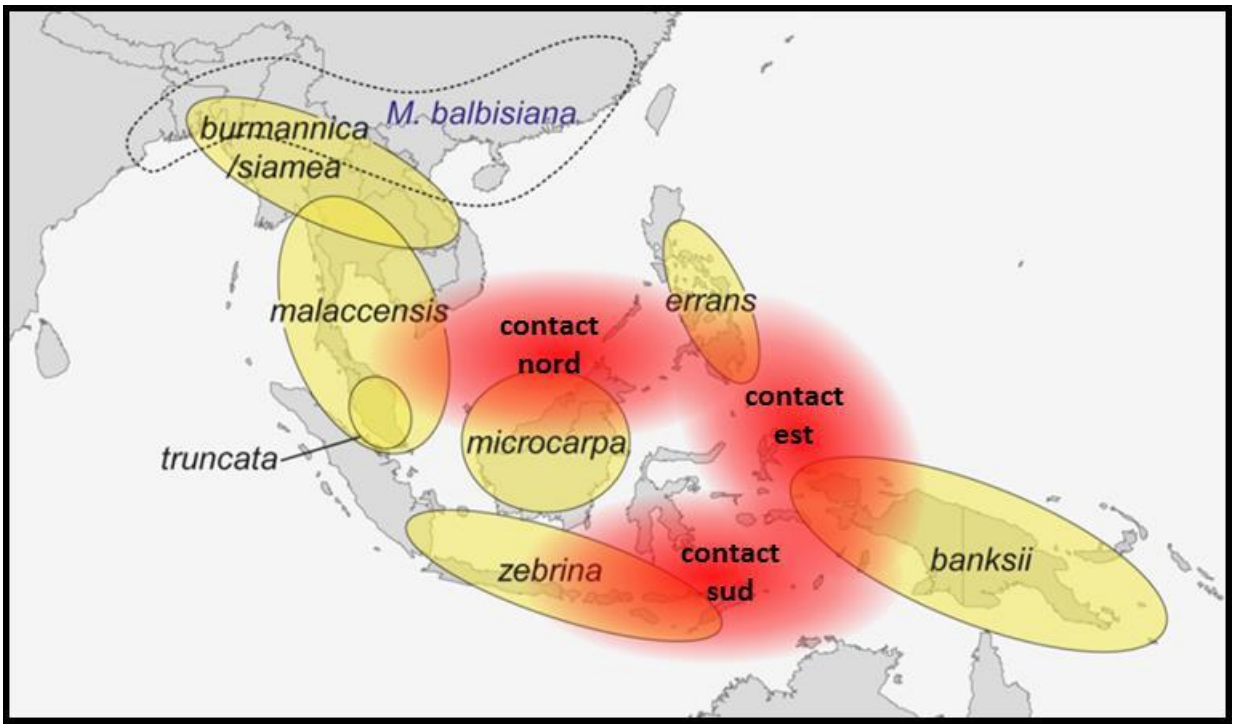


Figure 3 : Zones de contact entre les sous-espèces de *Musa acuminata* à l'origine des bananiers diploïdes cultivés. Les trois aires de contact principales ont permis le contact, au nord entre les sous-espèces *malaccensis*, *microcarpa* et *errans*, à l'est entre *errans* et *banksii* et au sud entre les sous-espèces *banksii*, *zebrina* et *microcarpa* (Figure modifiée de Perrier et al., 2011).

1.2 Origine des bananiers cultivés

Des études interdisciplinaires associant des données de marquage moléculaire (Carreel, 1994; Carreel et al., 1994, 2002; Hippolyte et al., 2012; Risterucci et al., 2009), d'archéologie et de linguistique ont permis récemment de préciser les hypothèses sur les origines des bananiers cultivés (Perrier et al., 2009, 2011). Les bananiers sauvages ont évolué en Asie du Sud et des populations isolées géographiquement par montée du niveau de la mer ont divergé en sous-espèces avec notamment l'apparition de réarrangements chromosomiques (Shepherd, 1999). La domestication des bananiers comprend une première étape d'hybridation entre diploïdes sauvages des sous-espèces de *Musa acuminata* (appelés AAw dans ce document). Ainsi, durant l'Holocène, des migrations humaines ont eu lieu en Asie du Sud-Est. Ces populations auraient emmené avec elles leurs bananiers remettant ainsi en contact les différentes sous-espèces et espèces, rendant possible ces évènements d'hybridation (Perrier et al., 2011). Trois zones d'hybridations ont été supposées dans les îles d'Asie du Sud-Est et la Mélanésie (**Figure 3)**. La zone de contact la plus au Nord (entre les Philippines et l'Asie continentale du Sud-Est) est à l'origine d'hybrides entre les sous-espèces *malaccensis*, *microcarpa* et *errans*. Cette zone est à l'origine d'un grand nombre de bananiers cultivés AA (dits AAcv dans ce document) qui ont contribué à un grand nombre de triploïdes AAA. La zone de contact Est est à l'origine d'hybrides entre les sous-espèces *errans* et *banksii* ainsi que *Musa balbisiana*. La zone de contact Sud a vu l'apparition d'hybrides entre les sous-espèces *banksii*, *zebrina* et *microcarpa* qui sont à l'origine d'un grand nombre de bananiers diploïdes AA et triploïdes AAA cultivés en Afrique. Lors d'une deuxième étape, les hybrides AA et AB se sont croisés entre eux ou avec des génotypes sauvages. Les hétérozygoties de structures chromosomiques des premiers hybrides ont entraîné la formation de gamètes diploïdes et donc la formation de génotypes triploïdes (Perrier et al., 2011). De Langhe et al., (2010) ont proposé que les cultivars résultent de plusieurs évènements de croisements et/ou rétrocroisements. Ensuite, la multiplication naturelle du bananier par voie végétative a favorisé la dissémination de quelques cultivars sur de grands territoires et l'apparition d'une diversité phénotypique, probablement par la fixation de variations somaclonales. Les clones issus de chacun de ces cultivars sont maintenant regroupés en sous-groupes (les plus importants sont les sous-groupes des Cavendish, des Gros Michel, des Lujugira-Mutika et des Plantains). Cette multiplication végétative a aussi favorisé la monoculture, ce qui a facilité l'adaptation des pathogènes (virus, bactéries, champignons, nématodes et insectes) et donc fragilisé la culture de la banane (Dita et al., 2010; de Lapeyre de Bellaire et al., 2010). Ces pratiques ont entraîné

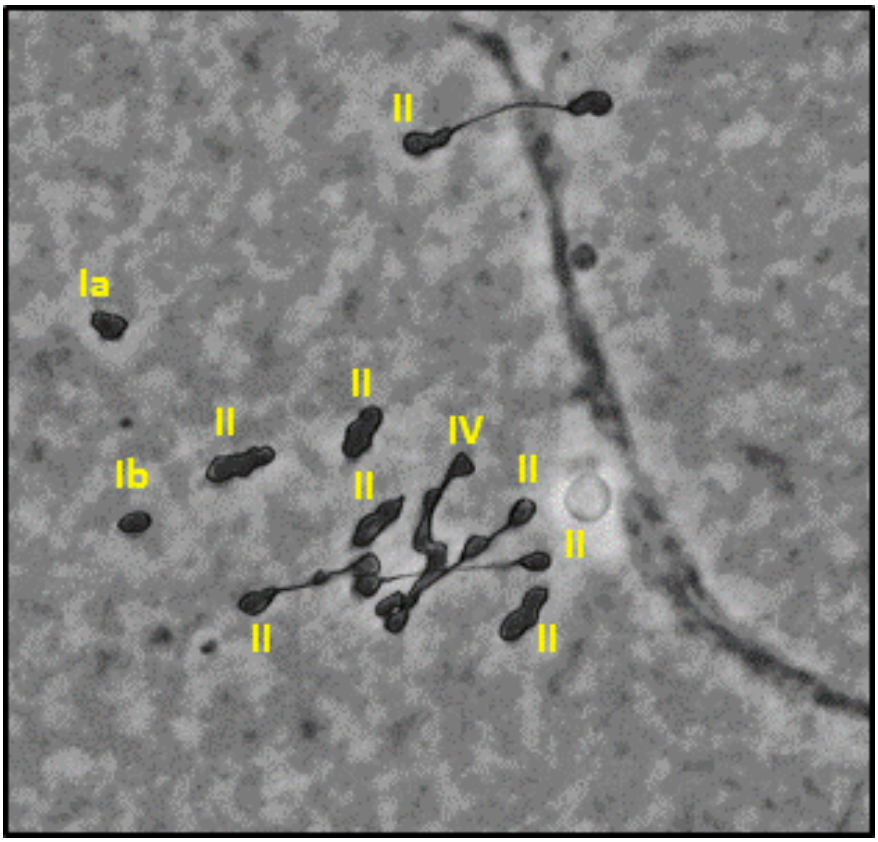


Figure 4 : Appariement de chromosomes de bananier au stade métaphase de la méiose. Huit paires de chromosomes sont appariées normalement sous forme de huit bivalents (II). Deux chromosomes ne sont pas appariés et forment deux monovalents (Ia et Ib). Quatre chromosomes forment un tetravalent (IV).

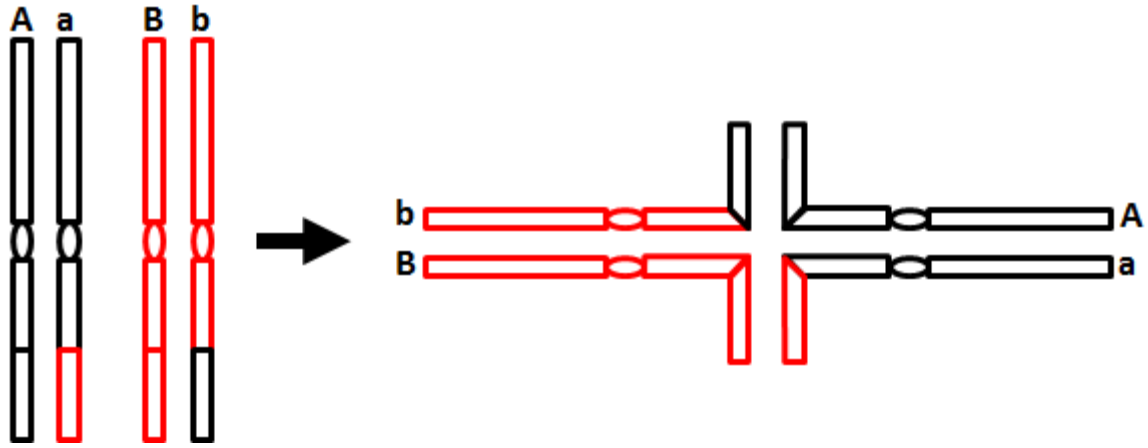


Figure 5 : Représentation schématique de l'appariement en tétravalent des chromosomes un individu hétérozygote de structure pour une translocation réciproque impliquant les extrémités de deux paires de chromosomes.

durant la première moitié du XX^{ème} siècle l'abandon mondial des cultivars du sous-groupe Gros Michel (AAA) et leur remplacement par les cultivars du sous-groupe Cavendish (AAA) suite à l'émergence de la maladie de Panama (Ploetz, 1994). La lutte contre ces maladies et ravageurs, quand elle est possible, requiert l'usage intensif de produits phytosanitaires qui affectent la viabilité économique des systèmes de production et qui a surtout un impact extrêmement négatif sur l'environnement et la santé des travailleurs agricoles. Dans ce contexte, il est urgent d'exploiter la diversité génétique des bananiers pour créer des variétés plus résistantes tout en produisant des fruits de qualité.

1.3 L'amélioration du bananier

Les programmes d'amélioration du bananier présentent la particularité que la finalité soit que la plante obtenue doit produire des fruits sans graines, donc être stérile ou à fertilité très réduite. Le niveau triploïde semble être le plus performant pour apporter cette fertilité très réduite et de bonnes performances agronomiques. Deux approches principales sont utilisées dans les programmes d'amélioration génétique du bananier (Bakry et al., 2009). La première approche consiste à croiser l'accession triploïde à améliorer avec une accession diploïde porteuse de caractères agronomiques intéressants. Ce croisement produit une descendance présentant différents niveaux de ploïdie avec de nombreux aneuploïdes. Dans cette descendance, des bananiers tétraploïdes sont directement sélectionnés pour être valorisés en plantation. Des bananiers tétraploïdes et diploïdes sont également sélectionnés pour être utilisés comme géniteurs dans des croisements 4x X 2x afin d'obtenir une majorité de bananiers triploïdes. La seconde approche vise à recréer des bananiers triploïdes à partir de bananiers diploïdes. Les bananiers triploïdes sont obtenus en croisant un bananier diploïde avec un bananier tétraploïde, obtenu par doublement chromosomique d'un diploïde. Cette seconde approche est en particulier utilisée par le programme d'amélioration génétique du CIRAD en Guadeloupe pour tenter d'obtenir des triploïdes proches du cultivar Cavendish quasi-infertile en utilisant comme parents les contributeurs diploïdes probables de ce cultivar révélés par les études de diversité moléculaire (Hippolyte et al., 2012; Raboin et al., 2005).
Historiquement, les bananiers utilisés par ces approches étaient issus de prospections dans leur zone d'origine et leur zone de culture. Pour assurer la fertilité et apporter des caractères de résistance aux différents pathogènes et ravageurs, des diploïdes sauvages ont été utilisés mais les hybrides obtenus portaient souvent de nombreux caractères rédhibitoires pour la qualité du fruit. Certains programmes d'amélioration génétique ont donc initié en amont un

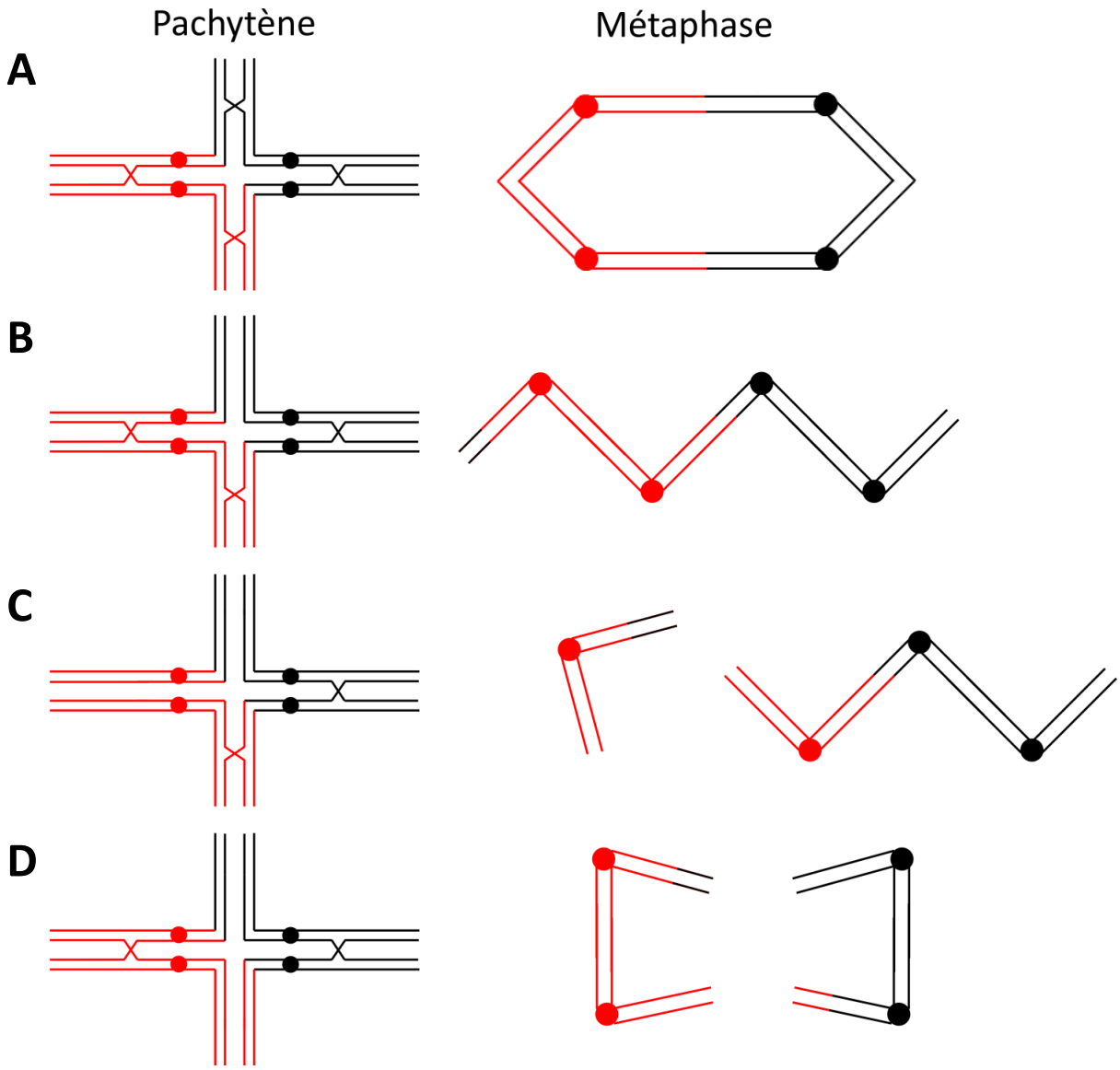


Figure 6 : Représentation schématique des structures observées en métaphase en fonction de la localisation des chiasma au stade pachytène (prophase) pour un individu présentant une translocation réciproque impliquant deux paires de chromosomes. Les deux chromatides sœurs sont représentées par deux barres parallèles, les points représentent les centromères et les régions homologues ont la même couleur. (**A**) Un anneau de quatre chromosomes est attendu en métaphase si des chiasma ont lieu dans les quatre segments. (**B**) Une chaîne de quatre chromosomes est attendue en métaphase si des chiasma ont lieu dans trois des quatre segments. (**C**) La présence de chiasma dans deux segments adjacents entraîne une structure à la métaphase comprenant une chaîne de trois chromosomes et un chromosome seul. (**D**) Des chiasma sur deux segments non adjacents entraînent une structure à la métaphase comprenant deux paires de chromosomes.

programme d'amélioration des géniteurs diploïdes. Toutefois, les descendances obtenues ont souvent un petit effectif et présentent de fortes distorsions de ségrégation. Dans ce contexte, il est particulièrement important d'en connaître les raisons afin d'orienter au mieux les croisements et les possibilités d'introgression des caractères. Des hypothèses de variations de structures chromosomiques entre les sous-espèces de *Musa acuminata* ont souvent été avancées pour expliquer cette perte de fertilité et les distorsions de ségrégation.

2- Appariements chromosomique chez les *Musa* impliqués dans les bananiers cultivés

Les appariements chromosomiques à la méiose ont été étudiés chez de nombreux bananiers séminifères, parthénocarpiques et hybrides intra *Musa acuminata* et inter *Musa acuminata* x *Musa balbisiana* (e.g. Dessauw, 1987; Dodds, 1943; Dodds et Simmonds, 1948; Fauré et al., 1993a; Hippolyte et al., 2010; Shepherd, 1999; Vilarinhos, 2004). Chez les bananiers séminifères, les méioses sont en général régulières avec la formation de 11 bivalents. Par contre, chez les bananiers parthénocarpiques et hybrides entre espèces et sous-espèces, des multivalents et univalents (**Figure 4**) sont fréquemment observés. Ces méioses perturbées ont été attribuées à la présence d'hétérozygoties de structures chromosomiques de type translocation, inversion, duplication (développées dans le point **6.1** de ce chapitre introductif). Ainsi, selon le type de figure d'appariement observé, on peut supposer différents types d'hétérozygoties structurales.

Par exemple, dans le cas d'une translocation réciproque impliquant deux paires de chromosomes, les régions homologues peuvent s'apparier en formant un groupe de quatre chromosomes (tétravalent) lors de la prophase I (**Figure 5**). La structure observée en métaphase dépend alors de la fréquence et de la localisation des chiasmas (Fauré et al., 1993a) (**Figure 6**). Lorsque des chiasmas ont lieu dans les quatre segments d'appariement, un anneau de quatre chromosomes est alors observable en métaphase (**Figure 6A**). L'apparition de chiasmas dans trois des quatre segments entraîne l'observation d'une chaîne de quatre chromosomes (**Figure 6B**). Si deux chiasmas se forment dans deux segments adjacents, on observera en métaphase une chaîne de trois chromosomes (trivalent) et un chromosome seul (dit monovalent ou univalent) (**Figure 6C**). Enfin, si deux chiasmas se forment dans deux segments opposés, deux groupes de deux chromosomes (2 paires de bivalents) seront alors observés (**Figure 6D**).

Dessauw (1987) conclut que tous les bananiers parthénocarpiques étudiés sont hétérozygotes de structure, alors que les bananiers séminifères étudiés sont homozygotes de structure.

Tableau 1: Groupes caryotypiques et nombre de translocations entre les différents groupes identifiés sur la base des figures d'appariement chromosomique à la méiose des *Musa acuminata* et *Musa balbisiana* (d'après Shepherd, 1999).

	group	Standard (ban, mic, mal)[a]	Northern Malayan (mal)[a]	Malayan highland (tru)[a]	Northern 1 (bur2, sia)[a]	Northern 2 (bur1, sia)[a]	Javanese (zeb)[a]	East Africa (ind)[a]
Musa acuminata	Northern Malayan	1						
	Malayan highland	3	4					
	Northern 1	2	3	*				
	Northern 2	3	4	*	1			
	Javanese	1	2	?	3	4		
	East Africa	1	2	?	3	4	2	
	Musa balbisiana	1	?	?	2	3	?	?

a : ban, *banksii* ; mic, *microcarpa* ; mal, *malaccensis* ; tru, *truncata* ; bur1, *burmannica* ; bur2, *burmannicoides* ; sia, *siamea* ; zeb, *zebrina* et ind, ssp indéterminés.

* différence structurale supposée mais non identifiée

? Pas de données

L'étude des méioses chez les hybrides dont on connaît les parents permet alors d'identifier les variations des structures chromosomiques entre les espèces et sous-espèces de *Musa*. Sur la base de ses propres travaux et de travaux antérieurs dont il fait la synthèse, Shepherd (1999) a défini au sein des *M. acuminata* sept groupes « caryotypiques » (ou groupes de translocation) à l'intérieur desquels les accessions seraient homozygotes de structure (Standard, Northern Malayan, Northern 1, Northern 2, Malayan Highland, Javanese et East African). Les regroupements correspondent en partie à la structuration en sous-espèces avec toutefois des sous-espèces subdivisées dans deux groupes (**Tableau 1**). Le groupe dit 'Standard' est le groupe le plus large et comprend les sous-espèces *microcarpa*, *banksii* et la plupart des accessions *malaccensis*. Le groupe 'Northern Malayan' regroupe des accessions rattachées à la sous-espèce *malaccensis*. Le groupe 'Northern1' comprend des accessions rattachées à la sous-espèce *burmannicoides* et des accessions appartenant à la sous-espèce *siamea*. 'Northern2' regroupe des accessions rattachées à la sous-espèce *burmannica* et des accessions appartenant à la sous-espèce *siamea*. Les groupes dits 'Malayan Highland' et 'Javanese' sont définis par les sous-espèces *truncata* et *zebrina* respectivement alors que le dernier groupe, 'East African', correspond à une accession dont la sous-espèce n'a pas été identifiée. Les hybrides intergroupes seraient hétérozygotes pour 1 à 4 translocations ou inversions (**Tableau 1**). Parmi les neuf évènements de translocation identifiés, seulement deux seraient partagés entre les groupes Northern1 et Northern2. L'étude des méioses des hybrides entre *Musa acuminata* et *Musa balbisiana* laisse, quant à elle, supposer la présence d'au moins une translocation entre le groupe dit 'Standard' de *M. acuminata* et *M. balbisiana* (Shepherd, 1999) (**Tableau 1**). Il est important de noter que cette synthèse comprend des études bien antérieures et des problèmes de correspondance entre certaines accessions étudiées à l'époque et leur version actuelle en collection se posent (Xavier Perrier com. pers.). Par ailleurs, cette synthèse porte sur un nombre relativement faible d'accessions par rapport à la diversité des bananiers.

Les génotypes qui présentent des hétérozygoties de structure chromosomique peuvent générer des gamètes dépourvus de certaines régions chromosomiques alors que d'autres régions peuvent se retrouver en plusieurs copies. En général, ces gamètes ne sont pas viables. La mortalité d'un certain nombre de gamètes haploïdes peut entraîner une augmentation artificielle de la proportion de gamètes diploïdes naturellement produits (Ramsey et Schemske, 1998) facilitant la formation de polyploïdes. D'autre part, cette mortalité entraîne en général une baisse de fertilité. Ainsi, la baisse de fertilité observée chez la plupart des bananiers parthénocarpiques a souvent été attribuée à cette réduction du nombre de gamètes

Tableau 2 : Synthèse des cartes génétiques publiées sur l'espèce Musa acuminata.

Croisement	Carte	Sous-espèce de Musa acuminata	Taille de la population	Nombre de marqueurs	Marqueurs distordus (%)	Nombre de groupes de liaisons	Groupes de liaisons les plus distordus	Informations cytogénétiques	Références
auto-fécondation	SFB5	hybride[1]	92	77	36%[a]	15	SFB5-LG12	2 translocations réciproques	Fauré et al., 1993b
auto-fécondation	M53	hybride[2]	89	185	26%[b]	18	M53-LG9, M53-LG7	-	Noyer et al., 1997
auto-fécondation	CAM	hybride[3]	154	120	59%[a]	14	CAM-LG1, CAM-LG2, CAM-LG5	2 translocations non réciproques	Vilarinhos, 2004
TMB_{2x} 6142-1 x TMB_{2x} 8075-7	TMB_{2x} 6142-1	hybride[4]	81	231	14.5%[b]	15	A11, A12, A14	-	Mbanjo et al., 2012
6142-1-S x TMB_{2x} 8075-7	6142-1-S	burmannica	58	152	8.2%[b]	16	B2	-	Mbanjo et al., 2012
multi parentale	TMB_{2x} 8075-7	hybride[5]	139	316	40.6%[b]	15	C1, C6, C7, C9	-	Mbanjo et al., 2012
Bornéo x Pisang lilin	Bornéo	microcarpa	180	261	12%[b]	11	chromosomes 6, 8	2 polymorphismes de structure	Hippolyte et al., 2010
Bornéo x Pisang lilin	Pisang lilin	malaccensis	180	359	24%[b]	11	chromosomes 1, 2, 3, 4, 10	1 variation structurale + 1 inversion	Hippolyte et al., 2010
auto-fécondation	Pahang	malaccensis	180	652	17%[b]	11	chromosomes 1, 4	-	D'Hont et al., 2012

[a]: déviation significative par rapport à une ségrégation mendelienne $\chi 2$ $p < 0.05$

[b]: déviation significative par rapport à une ségrégation mendelienne $\chi 2$ $p < 0.005$

[1]: génotype hybride comprenant une composante *banksii*

[2]: génotype hybride comprenant une composante *malaccensis* et *banksii*

[3]: génotype hybride comprenant une composante *burmannicoides* et *banksii*

[4]: génotype hybride comprenant une composante *burmannica*

[5]: génotype hybride comprenant une composante *burmannicoides*

viables à cause de variations structurales (Dessauw, 1987; Shepherd, 1999). Cependant, Fauré et al., (1993a) nuancent cette conclusion par l'observation d'accessions très fertiles et pourtant identifiées comme étant hétérozygotes de structure sur la base d'observations cytologiques. Suite à ces observations, Fauré et al., (1993a) reprennent les conclusions de Dodds et Simmonds, (1948) qui suggèrent que la baisse de fertilité des bananiers n'est pas exclusivement due aux hétérozygoties de structures mais que des gènes peuvent également être impliqués.

3- Cartographie génétique et distorsions de ségrégation chez les bananiers

A ce jour, relativement peu de cartes génétiques ont été réalisées chez le bananier. L'obtention de descendances de taille significative est souvent difficile chez cette espèce et passe par une étape assez délicate de sauvetage d'embryon par culture *in-vitro* (Bakry, 2008). D'autre part, lors du maintien des populations au champ, un taux élevé de mortalité est souvent observé.

3.1 Les cartes génétiques de l'espèce Musa acuminata

Jusqu'à récemment, seule l'espèce *Musa acuminata* était étudiée dans des travaux de cartographie génétique. Les informations relatives à ces cartes sont synthétisées dans le **Tableau 2**.

La première carte génétique (SFB5) produite chez le bananier (Fauré et al., 1993b) a été réalisée à partir d'une population F2 de 92 individus, issue de l'autofécondation d'un individu F1 provenant du croisement entre un bananier parthénocarpique appartenant à l'espèce *Musa acuminata* (AAcv) 'SF265' et une accession séminifère (AAw) 'Banksii' appartenant à la sous-espèce *banksii*. Cette carte très partielle, comprenant 77 marqueurs (52 RFLP, Restriction Fragment Lenght Polymorphisms, 1 isozyme et 24 RAPD, Random Amplified Polymorphic DNA) regroupés en 15 groupes de liaison, a révélé une forte proportion de distorsions de ségrégation. Ainsi, 36% des marqueurs présentent une ségrégation non conforme avec une ségrégation mendélienne ($p<0.05$). La répartition des marqueurs distordus sur la carte est aussi fortement biaisée avec un groupe de liaison (SFB5-LG12) présentant 8 marqueurs, tous fortement distordus. Le faible nombre de marqueurs polymorphes disponibles à l'époque chez le bananier ne permettait pas de tirer de conclusion sur la répartition des marqueurs distordus pour les autres groupes. L'étude de l'appariement chromosomique à la

méiose du parent de la population a révélé la formation de multivalents (univalents, trivalents, tétravalents et hexavalents) qui ont été interprétés comme résultant de la présence de deux translocations réciproques à l'état hétérozygote impliquant trois paires de chromosomes (Fauré, 1993).

Une seconde carte (M53) a été construite à partir de 89 individus issus d'une autofécondation du cultivar diploïde AAcv 'M53' (Noyer et al., 1997). Ce cultivar est un hybride synthétique [*M. acuminata spp. malaccensis* 'Kedah' x *M. acuminata spp. banksii* 'Samoa'] x [AAcv 'Paka' x *M. acuminata spp. banksii* 'Samoa']. Cette carte génétique comprend 185 marqueurs (120 AFLP, Amplification Fragment Length Polymorphism, 40 RFLP, 24 SSR, Simple Sequence Repeat, et 1 isozyme) distribués sur 18 groupes de liaison. La carte M53 comprend un taux de marqueurs distordus de 26% ($p<0.005$) localisés majoritairement sur deux groupes de liaisons (Jean-Louis Noyer com. pers.).

Une troisième carte (AFCAM20) a été réalisée à partir d'une population F2 de 154 individus issue de l'autofécondation d'un hybride F1 'CAM20' entre *M. acuminata.* ssp *burmannicoides* 'Calcutta4' et *M. acuminata.* ssp *banksii* 'Madang'. Cette carte qui comprend 120 marqueurs (81 AFLP, 20 RFLP et 19 SSR) groupés en 14 groupes de liaison comprend une très forte proportion de marqueurs distordus (59%, $p<0.05$) (Vilarinhos, 2004). Avec une telle proportion de marqueurs distordus, tous les groupes de liaison sont affectés à l'exception du groupe CAM-LG7. On peut noter une concentration particulièrement importante de marqueurs distordus dans les groupes CAM-LG1, CAM-LG2 et CAM-LG5.
L'analyse de la méiose du parent F1 'CAM20', a mis en évidence la présence de figures d'appariement en multivalents (univalents, trivalents, tétravalents et pentavalents) qui a été interprétée comme résultant de la présence d'hétozygoties structurales en la forme d'au moins deux translocations non réciproques (Vilarinhos, 2004). Des indices supplémentaires de la présence de ces translocations ont été apportés par l'étude plus précise du groupe CAM-LG2 par la technique du BAC-FISH (Vilarinhos, 2004). Cette étude n'a pas permis de conclure sur la nature exacte des hétérozygoties structurales présentes chez cet individu. Par contre, elle montre clairement une différence de structure au sein des *M. acuminata* avec la localisation de deux clones BAC situés sur la même paire de chromosomes pour 'Calcutta4' et sur deux paires différentes de chromosomes pour 'Madang'.

Les cartographies génétiques suivantes ont été réalisées à partir de 180 individus issus d'un croisement bi-parental (BorLi) entre *M. a.* ssp *microcarpa* 'Borneo' (AAw) et un AAcv 'Pisang lilin' apparenté à la sous-espèce *malaccensis*. Deux cartes parentales et une carte consensus de référence ont été construites avec près de 500 marqueurs (167 SSR et 322 DArT, Diversity Arrays Technology) répartis sur 11 groupes de liaison (Hippolyte et al., 2010). Plus de 20% des marqueurs présentent des distorsions de ségrégation ($p<0.005$) avec une hétérogénéité importante dans leur répartition entre les groupes de liaison des cartes parentales. La carte dérivée du parent 'Borneo' comprend 12% ($p<0.005$) de marqueurs distordus avec des concentrations importantes sur les groupes 8 et 6. La carte dérivée du parent 'Pisang lilin' contient 24% de marqueurs distordus ($p <0.005$) avec des concentrations importantes sur les groupes 1, 2, 3, 4 et 10.

L'étude des méioses du parent 'Borneo' a montré des appariements chromosomiques perturbés sur environ 30% des observations avec la présence d'univalents, trivalents, tétravalents, pentavalents et hexavalents. Shepherd, (1999) avait identifié cette accession comme homozygote de structure sur la base de travaux similaires. Cette contradiction suggère que ce ne sont sans doute pas les mêmes accessions qui ont été étudiées. Les appariements chromosomiques lors de la méiose du parent 'Pisang lilin' sont eux aussi perturbés sur 60% des observations, avec la présence d'univalents, trivalents, tétravalents et d'un pont. Sur cette base, les auteurs ont supposé l'existence de deux zones d'hétérozygotie de structure impliquant trois paires de chromosomes. La présence de distorsions de ségrégation dans une telle proportion a un impact très négatif sur la capacité des algorithmes de cartographie génétique à produire un ordre de marqueurs correct dans les groupes de liaison. Les auteurs ont donc analysé les liaisons génétiques entre marqueurs à l'aide de représentations graphiques arborées des distances entre marqueurs. Ceci a permis de construire un ordre des marqueurs basé sur une approche statistique par parcimonie.

L'interprétation des figures d'appariement à la méiose et une approche par simulation des distorsions de ségrégation produites par la présence de variations structurales suggérées par ces figures, ont permis de proposer la présence d'une duplication entre un segment des groupes de liaison 1 et une des deux versions du groupe de liaison 4, et d'une inversion sur le groupe 10 chez 'Pisang lilin' (Hippolyte et al., 2010).

Trois autres cartes ont été générées par Mbanjo et al., (2012) à partir de deux populations de demi-frères. Leur père commun est un hybride *Musa acuminata* (AAh) entre un bananier hybride (AAh) 'SH-3362' et le AAw 'Calcutta4' (*M. a.* ssp *burmannicoides*). Le parent

femelle de la première population (P1) est un hybride diploïde entre le triploïde AAA 'Nyamwihogora' appartenant au sous-groupe Lujugira-Mutika et le diploïde AAw 'Long Tavoy' (*M. a.* ssp *burmannica*). Le parent femelle de la seconde population (P2) est le diploïde 'Long Tavoy'. La carte du parent paternel dérivée des deux populations comprend 40.6% de marqueurs distordus ($p<0.005$) avec les groupes de liaisons C1, C6, C7 et C9 particulièrement riches en marqueurs distordus. Au total, 14.5% et 8.2% des marqueurs sont distordus ($p<0.005$) pour le parent maternel des populations P1 et P2 respectivement. Ces deux cartes présentent également des groupes plus riches en marqueurs distordus, A11, A12 et A4 pour le parent maternel de la population P1 et B2 pour le parent maternel de la population P2. Des analyses de représentations graphiques arborées des distances entre marqueurs, similaires à celles utilisées par Hippolyte et al., (2010), ont permis aux auteurs d'émettre l'hypothèse de la présence d'hétérozygoties structurales sous la forme d'une translocation au niveau du groupe C10, et d'autres variations de structure sur les groupes A14 et B9.

Enfin, la carte génétique qui a servi à l'assemblage de la séquence de référence du génome A du bananier (D'Hont et al., 2012) est issue de l'autofécondation de 'Pahang' (AAw, *M. a.* ssp *malaccensis*). Cette carte est composée de 652 marqueurs (589 SSR et 63 DArT) dont 17% sont distordus ($p<0.005$). Les distorsions sont localisées très majoritairement dans les groupes de liaison 1 et 4 comme dans la carte de 'Pisang Lilin'. Depuis la carte BorLi, la nomenclature des chromosomes et des groupes de liaison est réconciliée ce qui permet de localiser les distorsions de ségrégation sur les chromosomes 1 et 4 pour les accessions 'Pahang', bananier séminifère *M. a.* ssp *malaccensis*, et 'Pisang lilin', un AAcv apparenté à la sous-espèce *malaccensis*.

Toutes les cartes impliquant *Musa acuminata* ont révélé d'importantes distorsions de ségrégation même lorsque des accessions très fertiles sont impliquées ('Bornéo', 'Pahang'). L'ensemble de ces cartes regroupe des accessions appartenant aux sous-espèces *M .a. banksii, M. a. microcarpa, M.a. malaccensis, M.a burmannica* et *M.a. burmannicoides* et les distorsions de ségrégation observées ont été attribuées à des hétérozygoties de structure chromosomique par leur auteurs. Pour quatre des cartes générées, les hypothèses d'hétérozygotie de structure sont appuyées par des études d'appariements chromosomiques à la méiose chez les parents des populations concernées (**Tableau 2**).

3.2 Les cartes génétiques impliquant l'espèce Musa balbisiana

Plus récemment, deux cartes génétiques ont été construites à partir de croisements impliquant au moins pour partie l'espèce *Musa balbisiana.*

Une autofécondation du diploïde *M. balbisiana* 'Pisang Klutuk Wulung' (PKW) a été analysée. Il s'agit de la seule carte impliquant uniquement l'espèce *M. balbisiana*. Cette carte a été générée dans le cadre du projet en cours du séquençage du génome B (Jin et al., in prep.). Dans cette population, 7.9% des marqueurs sont distordus ($p<0.005$) et sont distribués sur l'ensemble des groupes de liaison.

Une deuxième population de 184 individus issue du croisement entre le tétraploïde issu d'hybridation AAAB 'CRBP39' et le diploïde AA 'Pahang-Cameroun' a été analysé (Noumbissié et al., submitted). L'analyse de la fréquence de 105 allèles de microsatellites du parent tétraploïde CRBP39 dans la descendance montre une faible proportion de distorsions (4.8% $p<0.005$). Par contre, dans cette population, la moitié des descendants présente, pour certains marqueurs, trois allèles ou un seul allèle transmis par le parent femelle tétraploïde au lieu de deux. L'analyse fine de l'origine spécifique des marqueurs, de leur répartition dans les différents groupes de liaison et des corrélations entre marqueurs a permis de suggérer la présence d'une variation structurale impliquant les chromosomes 1 et 3 entre *M. acuminata* et *M. balbisiana*. Il est intéressant de noter que cette hétérozygotie de structure engendre des distorsions de ségrégation plus faibles que celles observées intra *Musa acuminata*. Dans le cas d'un parent tétraploïde, on peut penser que le passage par des gamètes « diploïdes » permet de compenser au moins partiellement les effets de la résolution non symétrique des multivalents qui entraîne la formation de gamètes déséquilibrés (avec des régions dupliquées et/ou manquantes). Les gamètes qui auraient été non viables à l'état haploïde du fait de l'absence de certaines zones géniques se retrouvent alors compensés par l'autre copie résultant de la diploïdie du gamète et restent donc viables.

Cette variation de structure est maintenant confirmée par le projet de séquençage en cours du génome B qui révèle la présence d'une translocation réciproque entre les chromosomes 1 et 3 entre les génomes A et B (Jin et al., in prep.).

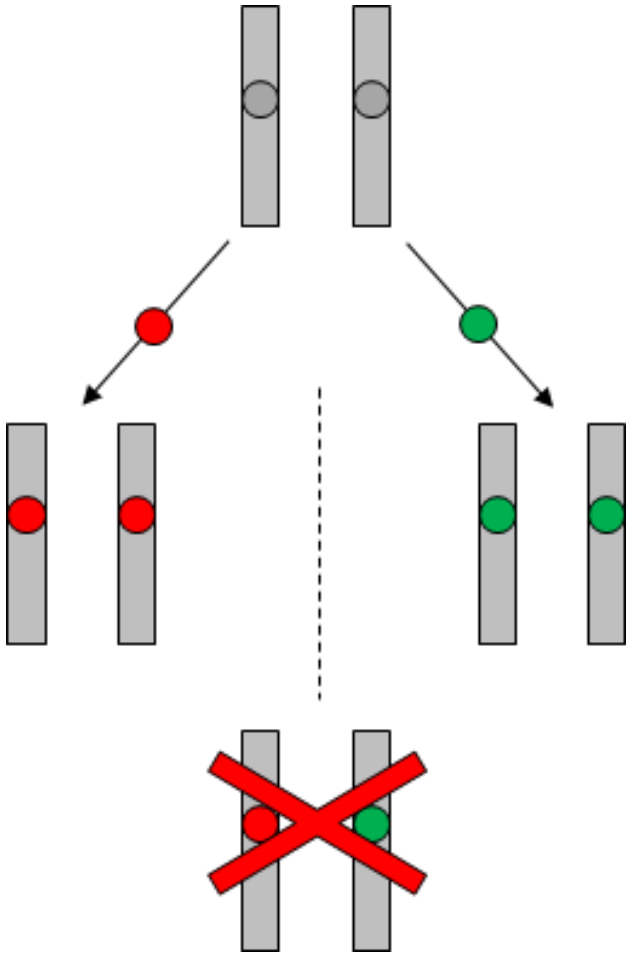


Figure 7 : Modèle d'incompatibilité hybride de Dobzhansky-Muller pour un seul locus. Une population ancestrale est séparée en deux populations isolées géographiquement qui divergent génétiquement à un locus et fixent un allèle différent (rouge ou vert). Lorsque ces espèces sont remises en contact, les hybrides F1 contiennent une copie de chaque allèle (rouge et vert) qui sont incompatibles et ce qui entraîne une baisse de fitness du zygote. Selon le rôle des gènes impactés, cette baisse de fitness peut impacter la survie du zygote ou sa fertilité.

4. Distorsions de ségrégation chez les plantes

L'analyse des liaisons entre marqueurs se base sur le postulat que les marqueurs étudiés ségrégent dans la descendance de façon mendélienne. Lorsque les fréquences alléliques observées diffèrent de celles attendues, on parle de distorsions de ségrégation. Des distorsions de ségrégation sont fréquemment observées chez les plantes. En plus des exemples concernant les bananiers évoqués plus haut, on peut citer le maïs qui fut la première plante où des distorsions ont été observées (Mangelsdorf et Jones, 1926), le blé (Takumi et al., 2013), les agrumes (Torres et al., 1985), *Arabidopsis* (Bikard et al., 2009), l'orge (Zivy et al., 1992), *Mimulus guttatus* (Fishman et al., 2001), le café (Lashermes et al., 2001), le riz (McCouch et al., 1988), *Medicago* (Jenczewski et al., 1997) ou encore le coton (Li et al., 2007).

4.1 Origine des distorsions de ségrégation

L'origine de ces distorsions de ségrégations peut être variée. On parle de sélection gamétique quand ces distorsions trouvent leur source dans une plus faible viabilité/transmission de certains gamètes et on parle de sélection zygotique lorsque les distorsions observées résultent d'une différence de viabilité entre les génotypes des différents hybrides générés. Les mécanismes à l'origine de ces différences de viabilité des gamètes ou des zygotes sont souvent difficiles à identifier. Différentes situations ont pu être caractérisées parmi lesquelles on peut citer:

(i) les combinaisons de gènes incompatibles : Ce mécanisme se base sur le modèle Dobzhansky-Muller (Dobzhansky, 1937; Muller, 1942). Ce modèle pose l'hypothèse qu'une mutation apportant un avantage survient dans une population et voit sa fréquence s'accroître dans la population jusqu'à sa fixation. Cependant, la "fitness" de l'allèle portant cette mutation n'est assurée que dans son propre contexte génétique mais peut être incompatible avec d'autres allèles présents chez d'autres génotypes. Le résultat est une stérilité ou perte de viabilité des hybrides portant ces combinaisons d'allèles. La **figure 7** schématise ce modèle pour un seul locus, comme observé chez le riz où deux gènes adjacents ont évolué séparément chez *Oryza sativa* ssp *indica* et *Oryza sativa* ssp *japonica*. Lorsque ces combinaisons alléliques sont mises en présence lors de la formation d'un hybride, cet hybride produit des gamètes (pollens) dont une partie est stérile par un effet d'interaction incompatible

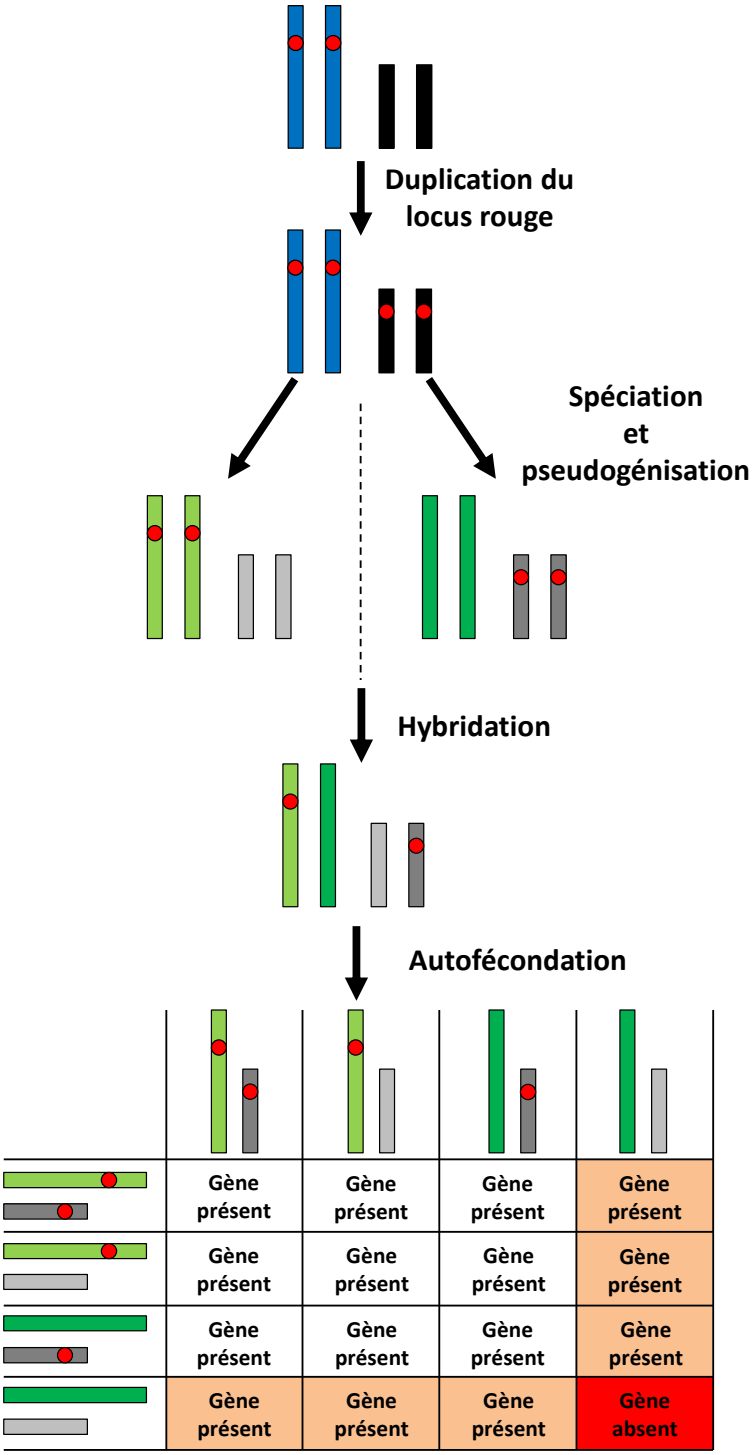


Figure 8 : Modèle d'évolution d'un gène dupliqué chez deux espèces divergentes et impact sur la fitness de l'hybride. Chez une population ayant deux chromosomes (bleu et noir), un gène (point rouge) est dupliqué entre le chromosome bleu et le chromosome noir. Suite à un mécanisme d'isolation, la population s'est scindée en deux populations qui ont divergées, avec en particulier la perte d'une copie du gène dupliqué chez les deux espèces mais chacune à partir d'un chromosome différent. La formation d'un hybride entre ces deux espèces et son autofécondation entraîne la formation de zygotes dont une partie est déficiente pour le gène anciennement dupliqué. L'absence de ce gène peut impacter la survie de ce type de zygote (case rouge). Par ailleurs, si ce gène a un rôle important dans la formation des gamètes, un certain nombre de zygotes (cases oranges et rouge) seront observés avec des fréquences inférieures à celle attendues.

entre les protéines produites par les allèles *indica* et *japonica* dans les microspores (revue dans Sweigart et Willis, 2012).

(ii) les différences de compétitivité entre génotypes : Fishman et al., (2001) émet l'hypothèse que les distorsions de ségrégation observées chez une F2 de *Mimulus guttatus x Mimulus nasutus* seraient liées à une meilleur compétitivité des produits d'un locus de *M. guttatus* par rapport à ceux de *M. nasutus*. Le locus concerné serait impliqué dans la croissance du tube pollinique qui serait plus compétitif lorsque les allèles *M. guttatus* sont présents. Ainsi, un des variants du locus de *M. guttatus* impliqué dans la croissance du tube pollinique se retrouve plus souvent dans la descendance qu'attendu en cas de ségrégation mendélienne.

(iii) la sélection préférentielle pendant la méiose (Meiotic drive) : Chez le maïs, un homologue du chromosome 10, dit 'abnormal 10' (ab10), migre préférentiellement vers la position de la mégaspore à l'origine de l'oosphère au cours de la méiose, d'où une augmentation de sa fréquence dans la descendance (Birchler et al., 2003; Buckler et al., 1999). Les allèles de ce chromosome sont alors en fréquence plus élevée dans la descendance que les allèles de son homologue, ce qui entraîne des distorsions de ségrégation sur ce chromosome.

(iv) les pertes de gènes : Une perte de gènes peut entraîner une perte de fonction. L'information génétique d'un locus donné est redondante à la fois car l'information est présente dans les deux copies d'un locus chez un diploïde, mais aussi souvent par la duplication des loci soit en tandem sur le même chromosome soit sur des positions éloignées dans le génome, on parle alors de gènes paralogues. Le devenir de ces gènes dupliqués au cours de l'évolution peut être variable. Cañestro et al., (2013) résument le devenir de ces copies en identifiant quatre possibilités:

- La première possibilité est la conservation des copies et de leur fonction.
- La seconde possibilité dite de sous-fonctionnalisation des copies consiste en un « partage » des tâches effectuées par la version ancestrale (unique) entre les différentes copies de gènes.
-La troisième possibilité est qu'une des copies dupliquée diverge pour acquérir une nouvelle fonction. On parle alors de néo-fonctionnalisation.
- La dernière possibilité est la perte de fonctionnalité d'une des copies suite à diverses mutations. On parle alors de pseudogénisation. Dans ce contexte de pseudogénisation de gènes dupliqués, il est alors facile d'envisager des pertes différentielles de gènes dupliqués

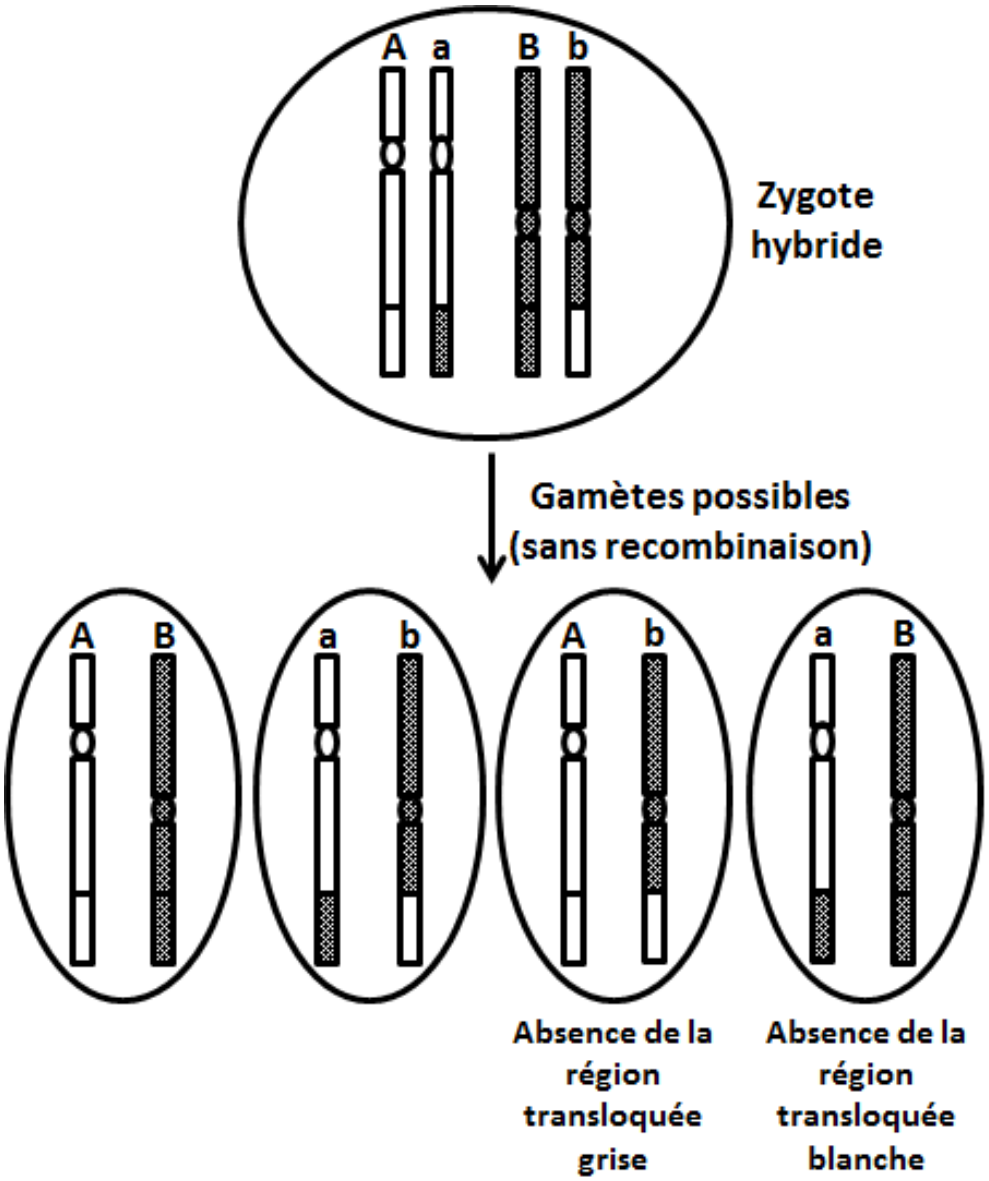


Figure 9 : Représentation schématique des gamètes possibles chez un hétérozygote de structure pour une translocation réciproque entre les chromosomes A et B. Sur les quatre types de gamètes possibles, deux gamètes sont déficients pour la région distale du chromosome A ou la région distale du chromosome B respectivement. Des gènes sont donc manquants, ce qui entraîne une baisse voir une perte de viabilité du gamète.

chez deux espèces divergentes. La réunion de ces deux génomes divergents sous la forme d'un hybride peut entraîner des baisses ou des absences de viabilité dans les gamètes ou la descendance de ces hybrides (**Figure 8**).

Bikard et al., (2009) ont mis en évidence une mortalité de certains embryons issus de l'autofécondation d'hybrides de deux lignées d'*Arabidopsis thaliana*. Dans l'une de ces deux lignées, un gène est dupliqué et présent sur les chromosomes 1 et 5 et la copie sur le chromosome 1 n'est plus exprimée. Dans l'autre lignée, ce gène n'est pas dupliqué et n'est présent que sur le chromosome 1. Ainsi dans la descendance, les individus homozygotes pour la copie non exprimée du chromosome 1 (issue de la première lignée) et l'absence du gène sur le chromosome 5 (issue de la seconde lignée) se retrouvent sans expression de ce gène ce qui entraîne la mortalité de l'embryon.

Les variations de structure chromosomique peuvent aussi entraîner des pertes de régions chromosomiques et donc de gènes. L'hétérozygotie structurale chez un hybride, en perturbant sa méiose avec la formation de multivalents et univalents, modifie les ségrégations chromosomiques. Ainsi, lors de l'anaphase, selon le mode de résolution des multivalents, leur répartition dans les cellules à l'origine des gamètes ainsi que celle des univalents, les gamètes peuvent se retrouver avec des déficiences et/ou des duplications de segments chromosomiques (**Figure 9**). La viabilité de ces gamètes est en général faible ou absente (Shepherd, 1999). Cependant Fauré (1993) suppose que la fertilité de ces gamètes dépendrait de la taille de la zone dupliquée et/ou déficiente ainsi que de la nature de la zone impliquée. Cette baisse ou perte de fertilité gamétique sera reflétée dans la descendance d'un hybride structural par une distorsion de ségrégation des marqueurs localisés sur les chromosomes impliqués dans le réarrangement.

Il existe assez peu de cas documentés chez les plantes. Jáuregui et al., (2001) ont montré que la présence d'une translocation réciproque impliquant les groupes de liaisons 6 et 8 entre *Prunus amygdalus* et *Prunus persica* entraînait chez un hybride une perturbation de sa méiose, une réduction de la fertilité mâle de moitié et des distorsions de ségrégation des marqueurs localisés dans les zones transloquées et à leur proximité. Dans un croisement interspécifique chez *Helianthus*, la baisse de viabilité des grains de pollen a également été attribuée à deux translocations réciproques présentes à l'état hétérozygote chez l'hybride (Quillet et al., 1995). Dans cette analyse, les marqueurs associés au caractère de viabilité pollinique sont distordus. Chez des hybrides interspécifiques de lentilles (*Lens*) la présence de marqueurs distordus a été associée à la présence d'une translocation réciproque à l'état hétérozygote (Tadmor et al., 1987).

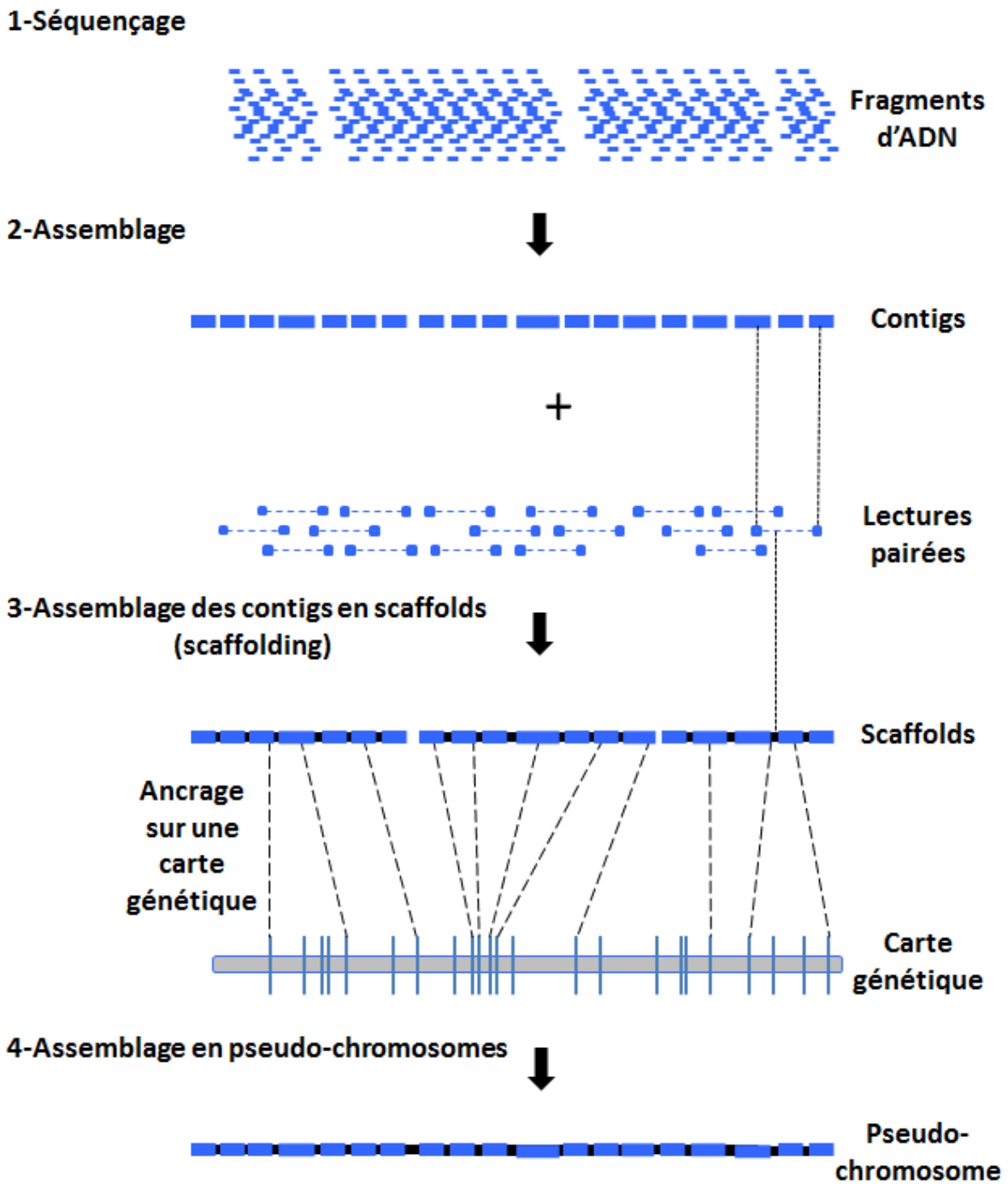


Figure 10 : Description des étapes principales conduisant à l'assemblage d'une séquence de référence en utilisant une approche de WGS. (1) L'ADN total d'une accession est séquencé. (2) Les lectures obtenues sont assemblées en contigs. (3) Les contigs sont ensuite assemblés en scaffolds en utilisant des lectures pairées (lectures obtenues par séquençage des extrémités de fragments d'ADN). (4) Les scaffolds obtenus sont ensuite groupés et ordonnés sur la base d'une carte génétique.

4.2 Conséquences sur les cartes génétiques

Les zones génomiques responsables de ces sélections gamétiques ou zygotiques sont les plus impactées par les distorsions de ségrégation. Les marqueurs les plus éloignés de ces zones seront moins, voire pas distordus du fait des recombinaisons qui randomisent la distribution des allèles et donc permettent de retrouver une distribution mendélienne. Par ailleurs, lorsque ces sélections impliquent des zones localisées sur des chromosomes différents (comme dans le cas iv) pertes de gènes), ces zones qui devraient ségréger indépendamment dans la descendance se retrouvent alors plus ou moins souvent liées génétiquement. Il y a alors une association génétique entre des marqueurs qui ne sont pas liés physiquement (=pseudo-liaisons). En cartographie, le résultat de ces associations entraîne le regroupement de ces marqueurs dans un seul groupe de liaison (Hippolyte et al., 2010; Jáuregui et al., 2001; Tadmor et al., 1987). De plus, dans le cas d'hétérozygotie structurale, l'apparition de ces pseudo-liaisons peut coïncider avec une réduction de la liaison entre les marqueurs situés de part et d'autre du point de réarrangement.

5-La séquence de référence : une nouvelle ressource pour l'analyse du génome des bananiers

Les deux premières espèces de plantes dont le génome a été complètement séquencé sont *Arabidopsis thaliana* (The Arabidopsis Genome Initiative, 2000) et le riz (Sequencing Project International Rice Genome, 2005). Ces deux génomes ont été séquencés en utilisant une approche dite de 'BAC par BAC' qui consiste à séquencer un set minimum de BAC chevauchants qui couvre l'ensemble du génome (revue par Feuillet et al., 2011). La sélection de ce set minimum de BAC peut se faire à l'aide d'une carte physique construite par des approches classiques de 'finger printing' (Chen et al., 2002) ou par des approches NGS (Next Generation Sequencing), comme le 'whole genome profiling' (Philippe et al., 2012). L'évolution des techniques de séquençage permet maintenant de séquencer un génome par des approches dites de 'Whole Genome Sequencing' (WGS). De nombreuses variantes en fonction de la ou des technique(s) de séquençage choisie(s) existent mais toute ces approches peuvent être décrites en quatre grandes étapes résumées sur la **Figure 10**: (i) le séquençage de l'ensemble du génome, (ii) l'assemblage des séquences obtenues en contigs, (iii) l'ordonnancement des contigs en scaffolds en utilisant des BAC-end ou des séquences pairées et (iv) l'ordonnancement des scaffolds en pseudo-molécules sur la base d'une carte génétique.

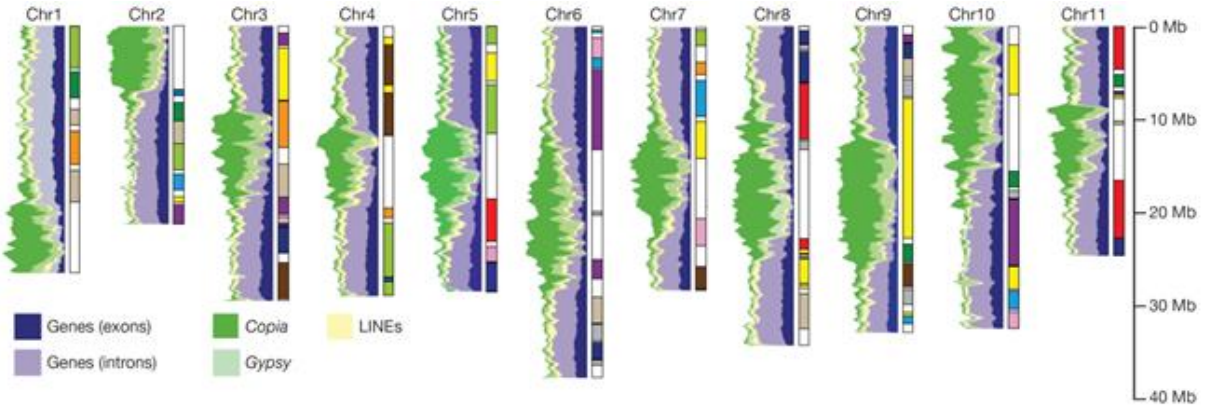


Figure 11 : Représentions schématique de la séquence de référence du bananier (*Musa acuminata*). Les densités en gènes, introns, rétrotransposons Gypsy, Copia et LINE sont représentées respectivement en bleu foncé, bleu clair, vert clair, vert et jaune. Les blocs de couleur verticaux indiquent les relations de paralogie entre les 11 chromosomes de la banane (extrait de D'Hont et al., 2012).

La séquence du génome du bananier a été réalisée en utilisant une approche WGS à partir d'un haploïde doublé de l'accession 'Pahang' de la sous espèce *M. acuminata* ssp *malaccensis* (D'Hont et al., 2012).

Un total de 24 425 contigs a été obtenu après assemblage d'un mélange de lectures 454 (16.9x) et de lectures Sanger (3.07x). Ces contigs ont été assemblés en 7 513 scaffolds en utilisant les lectures Sanger obtenues par séquençage des extrémités de fragments de 10 kb (2.94x) et des BAC-ends (0.13x). La N50 des scaffolds (taille minimale du scaffold au-dessus de laquelle 50% de l'assemblage peut être trouvé) est de 1.3Mb. Le plus grand scaffold de l'assemblage fait une taille de 12 Mb alors que le plus petit a une taille de 2 kb. La taille totale de l'assemblage représente 472 Mb soit 90% de la taille estimée du génome (523 Mb). L'assemblage comprend 17.3% de N (séquences inconnues).

Au total, 70% de l'assemblage, soit 264 scaffolds, a pu être ancré sur les 11 chromosomes du bananier en utilisant une carte génétique construite à partir de 589 marqueurs microsatellites en combinaison avec 63 marqueurs DArT. Les scaffolds qui n'ont pas pu être ancrés ont été regroupés dans une douzième pseudo-molécule dit chromosome Un_random.

Cette séquence est considérée dans les standards actuels comme "un high quality draft" (Chain et al., 2009).

Parmi les 11 chromosomes, deux sont acrocentriques (chromosomes 1 et 2) et un troisième, le chromosome 10, a probablement un centromère proche de l'extrémité du chromosome (**Figure 11**).

Un total de 36 542 gènes codant des protéines a été prédit par les programmes d'annotation automatique.

Les séquences répétées annotées représentent 44% de l'assemblage avec 55% localisées dans les scaffolds non ancrés. Des séquences répétées sont également retrouvées dans la partie non assemblée des séquences, ce qui signifie que la proportion de séquences répétées pour l'ensemble du génome est bien supérieure à 44%. La proportion de séquences répétées d'un génome est corrélée à la taille de son génome de base (Kejnovsky et al., 2012). Les séquences répétées annotées sont majoritairement localisées dans les zones péricentromériques, ce qui est classique pour les génomes de plantes. Par contre, on peut observer chez le bananier une transition très rapide entre les zones riches en séquences répétées et les zones riches en gènes qui correspondent aux régions distales des chromosomes, ce qui est caractéristique des petits génomes (D'hont, 2005; Jeridi et al., 2011). Les rétrotransposons à LTR représentent la proportion la plus importante des séquences répétées. Parmi les rétrotransposons à LTR, les éléments de type Copia sont plus abondants (25.7%) que les éléments de type Gypsy (11.6%).

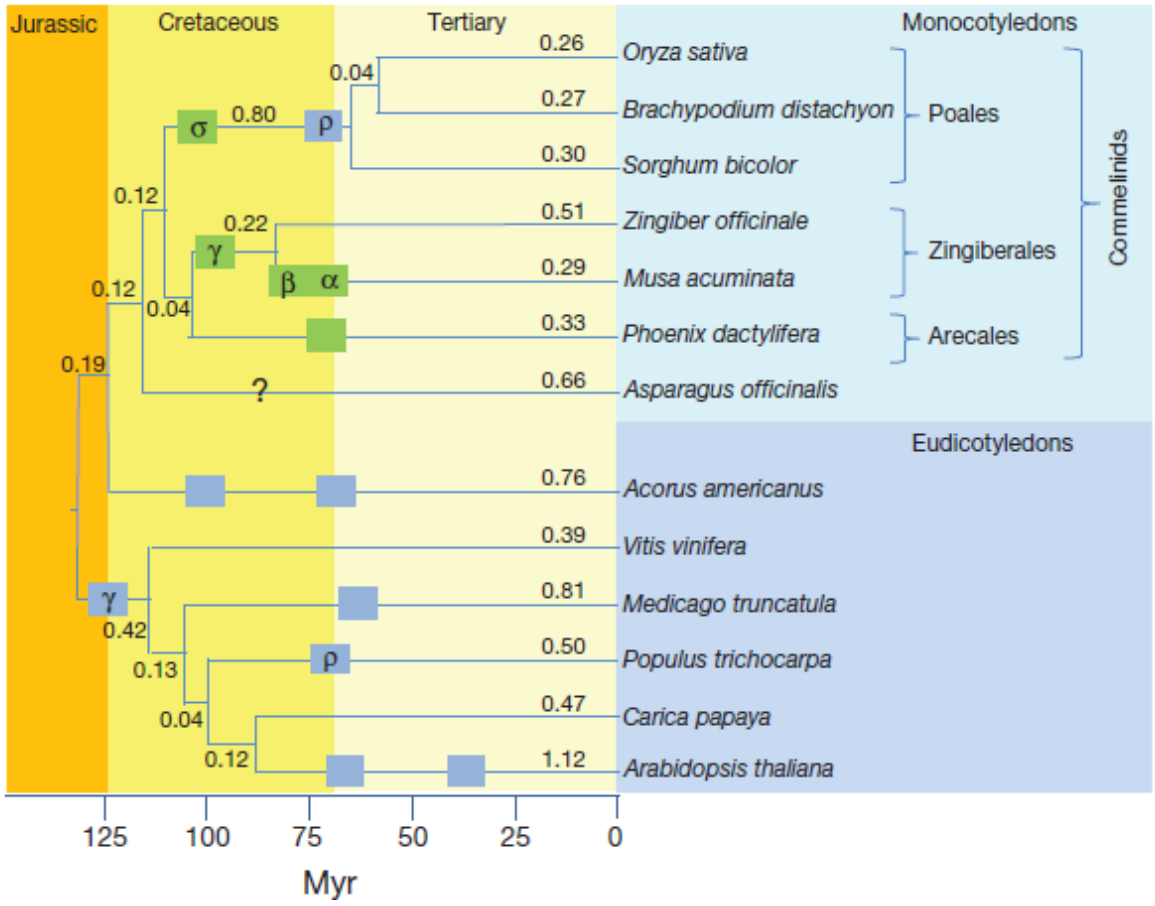


Figure 12 : Positionnement des évènements de duplication des génomes (*Whole Genome Duplication* : WGD) sur la phylogénie des monocotylédones et eudicotylédones. Les rectangles colorés indiquent les évènements de WGD identifiés. Les trois évènements spécifiques des Zingibérales identifiés grâce au séquençage du génome du bananier sont notés α, β et γ (extrait de D'Hont et al., 2012).

Il est intéressant de noter que chez le bananier, les éléments de type Copia sont les plus abondants alors qu'en en général ceux sont les éléments de type Gypsy qui prévalent (Hřibová et al., 2010). Les rétrotransposons de type LINE et les transposons ne représentent qu'une plus faible proportion du génome, soit respectivement, 5.5% et 1.3% du génome du bananier. Les éléments les plus grands chez le bananier sont les éléments de type Gypsy et Copia avec une taille moyenne respective de 5200 et 6600 bases.

Le séquençage des génomes a révélé de fréquents événements de duplication globale du génome (Whole Genome Duplication : WGD) au cours de l'histoire des plantes et l'importance de ces phénomènes dans l'évolution des génomes (Adams et Wendel, 2005; Cui et al., 2006; Jiao et Paterson, 2014; Soltis et al., 2014; Vanneste et al., 2014). Ainsi, l'analyse des segments dupliqués à l'échelle globale du génome du bananier a mis en évidence trois évènements anciens de WGD qui ont eu lieu après la divergence des Poales, Zingibérales et Arecales (**Figure 12**) (D'Hont et al., 2012). Un événement de WGD est daté à environ 100 millions d'années (Ma), ce qui indiquerait qu'il est commun aux Zingibérales. Les deux autres événements, plus récents, sont datés à environ 65 Ma.
Après un évènement de WGD, les gènes dupliqués peuvent évoluer selon les différents modes cités au paragraphe **4.1** de ce chapitre introductif. Cependant, la plupart des gènes dupliqués par WGD sont progressivement éliminés (= fractionation) (Cañestro et al., 2013). Chez le maïs (Schnable et al., 2011; Woodhouse et al., 2010) et *Arabidopsis* (Thomas et al., 2006), il a été montré que ces éliminations de gènes, quand elles font suite à un évènement de WGD, concernent plus particulièrement un des génomes ancestraux qui est également le moins exprimé. On parle alors de fractionation biaisée et de génome dominance. L'analyse de l'évolution des gènes après duplication a été réalisée chez le bananier ainsi que sur d'autres espèces séquencées. Cette étude a permis de proposer que selon le mode de WGD (allopolyploïdie ou autopolyploïdie) on observe ou non une fractionation biaisée et une dominance (Garsmeur et al., 2014). Chez le bananier, suite à la dernière WGD, on n'observe ni biais de fractionation ni dominance, ce qui permet de supposer une autopolyploïdie pour le dernier évènement de WGD.

La disponibilité d'une séquence de référence facilite les analyses portant sur l'évolution des familles de gènes du bananier puisque toutes les séquences des gènes d'une même famille sont alors disponibles et que la prise en compte des évènements de WGD permet une meilleure interprétation de l'évolution de ces familles de gènes (Jourda et al., 2014).

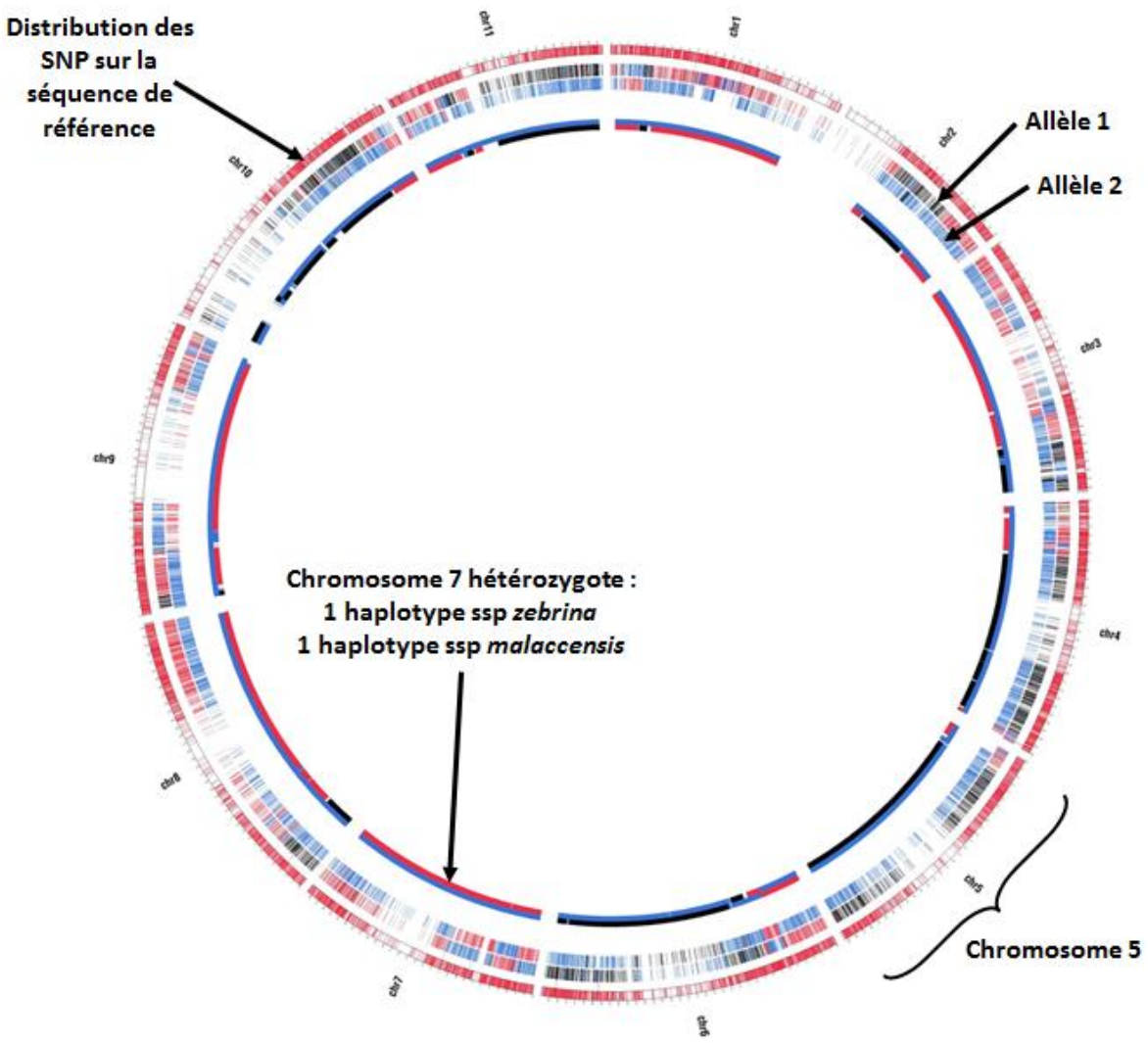


Figure 13 : Identification de la structure d'une accession de bananier par une approche de GBS. Le séquençage de plusieurs transcriptomes de bananiers a permis d'identifier un grand nombre de SNP qui ont été positionnés sur les 11 chromosomes du génome du bananier (traits rouges sur le cercle le plus extérieur). L'analyse de ces SNP a permis d'identifier un grand nombre d'allèles spécifiques à quatre sous-espèces de *Musa acuminata* (*banksii*, *zebrina*, *malaccensis* et *burmannica*). Ces allèles spécifiques sont utilisés pour étudier la composition du génome d'un bananier parthénocarpique. Les deux cercles médians représentent chacun, à un site donné, un des deux allèles spécifiques à une sous-espèce de *Musa acuminata*. Le codage des allèles en fonction de leur appartenance aux différentes sous-espèces est le suivant : *banksii*, vert (très peu présent dans cette accession) ; *zebrina*, rouge ; *malaccensis*, bleu et *burmannica*, noir. Une approche qui cherche à minimiser le nombre d'évènements de croisements (et donc de méioses/recombinaisons) est utilisée pour identifier les haplotypes présents dans le bananier étudié (cercle interne) (Com. pers. Nabila Yahiaoui, Cyril Jourda, Guillaume Martin).

La disponibilité d'une séquence de référence va également grandement faciliter les études de diversité ou de cartographies génétique basées sur du génotypage par séquençage (GBS) puisqu'il est alors plus facile de faire des recherches de SNP (Single Nucleotide Polymorphism).
Ainsi dans le cadre du projet ARCAD, plusieurs transcriptomes de bananier (*Musa acuminata*) séminifères et parthénocarpiques ont été séquencés, permettant l'identification de plusieurs milliers de SNP et l'attribution d'allèles spécifiques aux différentes sous-espèces de bananiers. Le positionnement de ces SNP sur la séquence de référence et l'analyse des allèles présents chez les bananiers parthénocarpiques permet d'étudier de façon plus précise la composition du génome des bananiers hybrides parthénocarpiques (**Figure 13**).

La disponibilité d'une séquence de référence permet également d'envisager de nouvelles approches pour détecter et mieux caractériser les variations de structures chromosomiques chez les bananiers.

6-La recherche de variations structurales

6.1 Les types de variations structurales

Dans la littérature, le terme de variations structurales englobe différents types de variations de la séquence génomique. Historiquement, les premières variations structurales identifiées ont été les réarrangements chromosomiques, c'est-à-dire des variations de taille importante impliquant des régions faisant plusieurs centaines de milliers de bases. Parmi ces variations, évoquées succinctement dans le point **2** de ce chapitre introductif, on identifie les duplications, les inversions et les translocations (**Figure 14**). Les **duplications** sont des variations structurales qui correspondent à des fragments génomiques présents en une seule copie dans un génome et qui sont présents en plusieurs copies dans un autre génome. Les **inversions** correspondent à des régions génomiques qui sont « retournées » chez un des deux génomes par rapport à l'ordre de l'autre génome. Enfin les **translocations** correspondent à des séquences qui se sont déplacées dans une autre région génomique chez un des génomes comparés. Cette translocation est dite **réciproque** si les deux régions s'échangent mutuellement un fragment. Elle est dite **non réciproque** dans le cas contraire. Les réarrangements chromosomiques représentent une question de recherche majeure depuis Dobzhansky et Sturtevant (1938). Ils ont été montrés comme des mécanismes importants pour

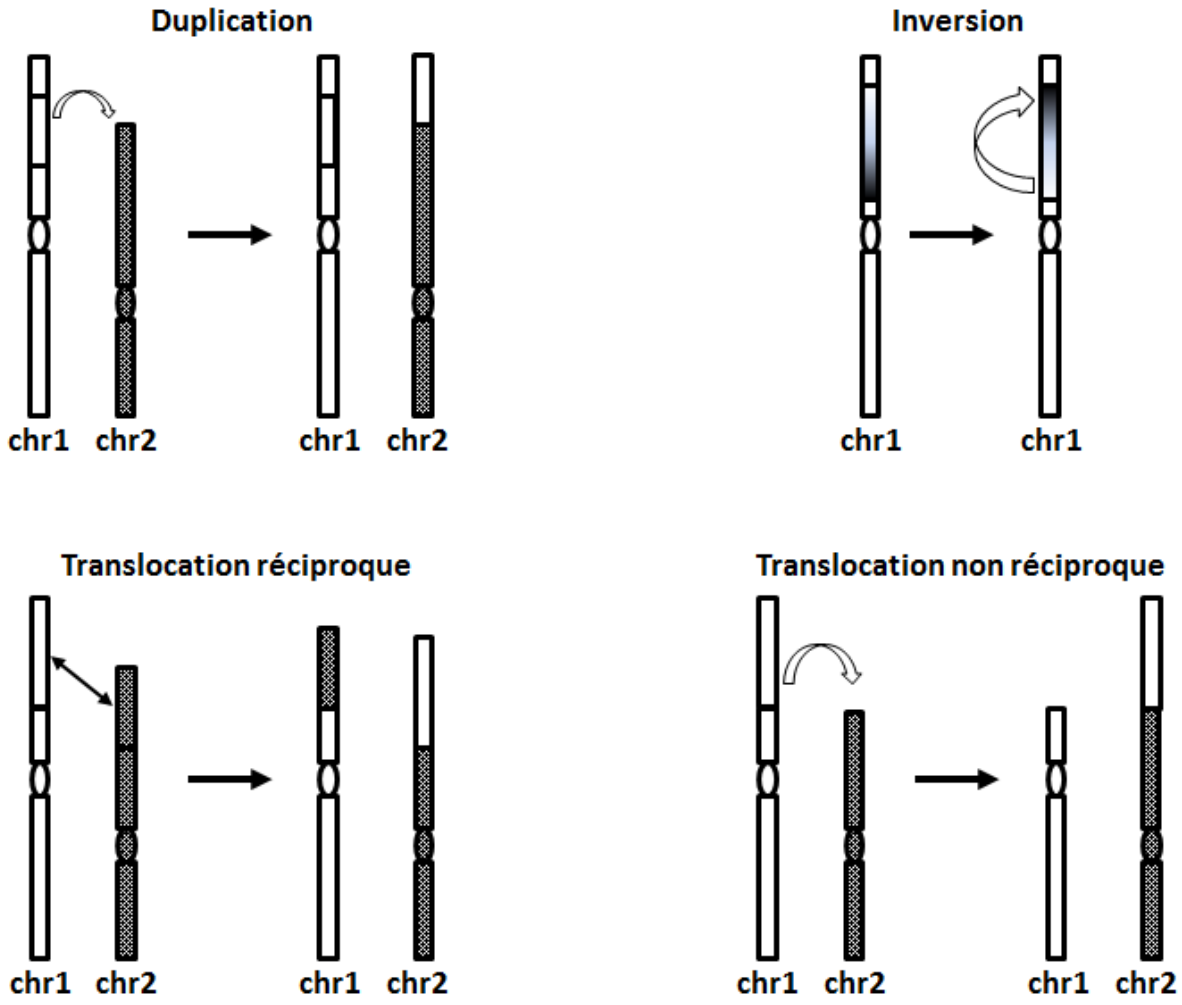


Figure 14 : Exemple de variations structurales de grande taille.

la spéciation et l'évolution (Rieseberg, 2001; White, 1962). Par contre, les mécanismes responsables de ces réarrangements sont toujours peu connus (Brown et O'Neill, 2009) et font actuellement l'objet de nombreuses recherches chez les plantes et les animaux (par exemple Carbone et al., 2009; Girirajan et al., 2009; Luo et al., 2009).

L'arrivée des séquences de génomes a permis, par des approches comparatives, d'étudier les variations de plus petite taille. Parmi ces variations, on identifie les très petites variations que sont les **SNP** et les **indels**. Les SNP sont les plus petites variations observées à l'échelle de la séquence de l'ADN qui correspondent au changement d'une base nucléotidique par une autre. Les insertions et délétions de petits fragments de séquence (<50 bp) sont regroupées sous le terme d'**indels**. Ces variations ne sont en général pas classées sous le terme de variation structurale.
La comparaison entre les génomes séquencés a d'autre part permis de mettre en évidence de nombreuses variations structurales de taille intermédiaire. Ces variations comprennent les différents types de variations structurales observés en cas de réarrangements chromosomique mais à plus petite échelle (inversions, duplications et translocations) ainsi que des insertions et délétions. La majeure partie de ces variations structurales est due aux mouvements des éléments transposables qui ont une taille entre quelques 100 pb et quelques kb et dont la position, le nombre et la séquence varient rapidement entre les génomes. Ainsi chez l'homme, les éléments Alu sont responsables d'une part importante des variations structurales observées entre génomes (Chen et al., 2009). Les éléments transposables peuvent également être impliqués dans les réarrangements chromosomiques de grandes tailles (Carbone et al., 2014; Gray, 2000; Xing et al., 2009). Parmi ces variations structurales de taille intermédiaire, on classe également les duplications de gènes (Cusack et Wolfe, 2007).

La classification de ces variations structurales en fonction de leur taille reste assez floue et aucune définition nette n'a encore émergé. Pour Alkan et al., (2011a), seuls les réarrangements de plus de 50 pb sont appelés variations structurales. Pour les mêmes auteurs, les variations structurales de grande taille font plusieurs centaines de kb alors que pour Korbel et al., (2007), les variations structurales de grande taille commencent à partir de 50 kb. Certains auteurs parlent de variations structurales de taille intermédiaire lorsque la taille de la variation est comprise entre 100 bp et 10 kb (Gong et al., 2013) alors que (Tuzun et al., 2005) considèrent une variation structurale de taille moyenne à partir de 8 kb.

Chez le bananier, les variations structurales suspectées sont capables d'entraîner des perturbations de la méiose et sont donc probablement de taille importante (plusieurs centaines de milliers de bases). On les qualifiera donc de variations structurales de grande taille.

6.2 Méthodes de détection des variations structurales

Différentes méthodes ont été utilisées et continuent d'être utilisées pour caractériser les variations structurales sans avoir recours à un génome de référence. Parmi elles, on peut citer :

(i) L'étude des figures d'appariements chromosomiques à la méiose :

Cette technique de cytogénétique dont le principe consiste à observer au microscope des figures d'appariement chromosomique à la méiose en métaphase et en anaphase et à interpréter les structures observées (univalents et multivalents) a été utilisée pour identifier des variations structurales chez le bananier mais également chez d'autres plantes comme le soja (Mahama et al., 1999), les lys (Xie et al., 2013) ou la gesse (Talukdar, 2010).

(ii) Fluorescence in situ Hybridization (FISH) :

Cette technique consiste à hybrider des fragments d'ADN (gènes, plasmides, BAC), marqués par fluorescence, directement sur les chromosomes en général au stade métaphase ou pachytène. Si ces sondes se localisent à différents endroits selon les génotypes, cela peut suggérer la présence de variations de structures chromosomiques entre ces génotypes. Cette technique a permis de mettre en évidence des variations du nombre de copies d'une région du chromosome 6 chez des tétraploïdes de pomme de terre (Iovene et al., 2013).

(iii) Comparaison de cartes génétiques :

Cette méthode consiste à comparer des cartes génétiques réalisées sur différents génotypes et à comparer les positions des marqueurs communs aux différentes cartes. Ce type d'approche a été utilisé pour comparer le blé et l'orge (Dubcovsky et al., 1996) et a permis de mettre en évidence une translocation et plusieurs inversions. Chez le bananier, ce type d'approche a permis de suggérer la présence de réarrangements chromosomiques ou hétérozygotie structurale (Hippolyte et al., 2010; Mbanjo et al., 2012; Vilarinhos, 2004).

La disponibilité de génomes séquencés ouvre l'accès à de nouvelles techniques pour identifier les variations structurales de grande taille. Parmi ces techniques, on peut citer :

Tableau 3 : Algorithmes de recherche de variations structurales (Extrait de Keane et al., 2014).

Algorithm	Description	Download	References
BreakDancer	Predicts del, ins, inv, and translocations using PEM.	http://gmt.genome.wustl.edu/breakdancer/current/	Chen et al., 2009
CNAseg	Identifies CNVs from NGS data.	http://www.compbio.group.cam.ac.uk/software.html	Ivakhno et al., 2010
cnD	HMM that uses read coverage to determine genomic copy number.	http://www.sanger.ac.uk/resources/software/cnd.html	Simpson et al., 2010
cn.MOPS	Mixture Of PoissonS Bayesian approach to detect CNVs.	http://www.bioinf.jku.at/software/cnmops	Klambauer et al., 2012
CNVer	Method that supplements the depth-of-coverage with PEM information, where mate pairs mapping discordantly to the reference serve to indicate the presence of variation.	http://compbio.cs.toronto.edu/cnver	Medvedev et al., 2010
CNVnator	Method for CNV discovery and genotyping from read-depth analysis of personal genome sequencing.	http://sv.gersteinlab.org/cnvnator	Abyzov et al., 2011
CNV-Seq	Method to detect CNV using shotgun sequencing.	http://tiger.dbs.nus.edu.sg/CNV-seq	Xie et Tammi, 2009
CREST	Clipping Reveals Structure, uses NGS reads with partial alignments to a ref. to map SVs at nucleotide level resolution.	http://www.stjuderesearch.org/site/lab/zhang	Wang et al., 2011
DELLY	Integrates paired-end and split-read analysis	www.korbel.embl.de/software.html	Rausch et al., 2012
Dindel	Bayesian method to call small indels by realigning reads to candidate haplotypes that represent alternative sequence to the reference, using a split-read approach.	http://www.sanger.ac.uk/resources/software/dindel	Albers et al., 2011
EWT	Event-wise testing, method based on significance testing. Error rate tested using the analysis of chromosome 1 from paired-end shotgun sequence data (30x) on 5 individuals	http://rdxplorer.sourceforge.net	Yoon et al., 2009
FREEC	Control-FREE Copy number caller that automatically normalizes and segments copy number profiles	http://bioinfo-out.curie.fr/projects/freec	Boeva et al., 2011
GASV-PRO	Combines both paired read and read depth signals into a probabilistic model for greater specificity	http://compbio.cs.brown.edu/software	Sindi et al., 2012
GenomeSTRiP	Genome STRucture In Populations, toolkit for discovering and genotyping structural variations using sequencing data. Twenty to thirty genomes required to get good results	http://www.broadinstitute.org/software/genomestrip/download-genome-strip	Handsaker et al., 2011
HYDRA	Localizes SV breakpoints by PEM. Uses a similar clustering strategy to VariationHunter. Accuracy evaluated using WGS slit-read mappings. Maps repetitive elements such as transposons and SD	http://code.google.com/p/hydra-sv	Quinlan et al., 2010
inGAP-sv	Scheme that uses abnormally mapped read pairs. Possible to distinguish HOM and HET variants.	http://ingap.sourceforge.net	Qi et Zhao, 2011
JointSLM	Allows to detect common CNVs among individuals using depth of coverage	http://www.mybiosoftware.com/population-genetics/11185	Magi et al., 2011

PEM, Paired-End Mapping; CNVs, Copy Number Variants; NGS, Next-Generation Sequencing; SVs, Structural Variants; SD, Segmental Duplication; WGS, Whole Genome Sequencing; pop., population; ind., individual; ref.,reference; seq., sequencing; ins, insertion; del, deletion; inv, inversion

(i) Chromosome painting :

Cette technique est une variante du FISH, dans laquelle les sondes ne sont plus constituées d'un petit fragment chromosomique mais d'un chromosome entier isolé par cytométrie de flux ou d'un ensemble de BAC couvrant un chromosome ou une portion de chromosome sélectionnée sur la base d'une séquence de référence. Chez les animaux, l'isolement de chromosomes sur la base de cytométrie en flux est beaucoup utilisée (e.g. Goureau et al., 1996; Rens et al., 2004; Shetty et al., 1999; Toder et al., 1997), il est beaucoup plus difficile à mettre en œuvre chez les plantes chez qui les chromosomes sont souvent peu différentiables par leur taille. Chez les plantes, une approche basé sur l'hybridation in situ de jeux de BACs correspondant à des chromosomes donnés sur la base d'une séquence de référence a été utilisée pour étudier les réarrangements structuraux qui ont eu lieu entre *Arabidopsis thaliana* et *Arabidopsis lyrata* (Lysak et al., 2006). Toutefois, cette méthode ne semble applicable qu'aux petits génomes pauvres en séquences répétées. Chez le bananier, il n'a actuellement pas été possible d'isoler les chromosomes par cytometrie en flux sur la base d'une différence de taille; par contre, il est envisageable de construire des sets de BAC correspondant à chacun des chromosomes pour tester le chromosome painting.

(ii) « Array painting » :

Cette méthode, décrite par Gribble et al., (2009), consiste à comparer la structure chromosomique d'un génotype connu avec la structure chromosomique d'un génotype inconnu. Pour cela, les chromosomes du génotype inconnu sont triés sur la base de leur taille par cytométrie en flux. Chaque chromosome de structure inconnue est ensuite marqué puis hybridé individuellement sur une puce (« Array ») contenant un assortiment ordonné de sondes dont on connaît la position sur le génome de référence. L'interprétation des hybridations, révélées par fluorescence, permet d'identifier les zones de rupture de la synténie due à des translocations ou duplications entre les deux génotypes comparés. De telles approches ont été utilisées, par exemple, pour comparer la structure chromosomique du gibbon à celle de l'homme (Gribble et al., 2004).

(iii) Comparaison de génomes complets à des cartes génétiques ou physiques :

La comparaison de cartes (génétiques ou physiques) à des séquences de génomes d'espèces apparentées permet aussi de détecter des variations structurales. La comparaison de la carte génétique de *Miscanthus sinensis* avec la séquence du sorgho a permis de mettre en évidence l'origine tétraploïde de *Miscanthus*, la fusion d'une des copies des chromosomes 4 et 7 en un

Tableau 3 (suite).

Algorithm	Description	Download	References
MindTheGap	Detect and assemble short and long insertions using k-mer based method	http://mindthegap.genouest.org	Rizk et al., 2014
MoDIL	Detection of small indels from clone-end sequencing with mixtures of distributions	http://compbio.cs.toronto.edu/modil	Lee et al., 2009
mrFast	Allows for the prediction of absolute copy-number variation of duplicated segments and genes	http://mrfast.sourceforge.net	Alkan et al., 2009
PEMer	Compatible with several NGS platforms. Simulation-based error models, yielding confidence-values for each SV	http://sv.gersteinlab.org/pemer	Korbel et al., 2009
Pindel	A pattern growth approach, to detect breakpoints of large deletions and medium-sized insertions from PEM reads	http://www.ebi.ac.uk	Ye et al., 2009
RetroSeq	Detects non-reference mobile elements such as LINE, SINE, and ERV.	https://github.com/tk2/RetroSeq	Keane et al., 2013
SoftSearch	Combines three analyses: split-read, read-pair, and single-end cluster.	http://bioinformaticstools.mayo.edu	Hart et al., 2013
SPANNER	SV detection for the pilot phase of the 1000 Genomes Project using low-coverage WGS of 179 ind. from 4 pop., high-coverage seq. of 2 mother-father-child trios, and exon targeted seq. of 697 ind. from 7 pop	https://github.com/chipstewart/Spanner	The 1000 Genomes Project Consortium, 2010
SplazerS	Method for split-read mapping, where a read may be interrupted by a gap in the read-to-reference alignment	http://www.seqan.de/projects	Emde et al., 2012
Splitread	Detects SV and indels from 1bp to 1Mb in exome data sets. Uses one end-anchored placements to cluster the mappings of subsequences of unanchored ends to identify size, content, and location	http://splitread.sourceforge.net	Karakoc et al., 2012
SRiC	Split-read identification, calibrated (SRiC).		Zhang et al., 2011
SVDetect	Identify discordant mate-pairs derived from NGS data produced by the Illumina GA and ABI SOLiD platforms	http://svdetect.sourceforge.net	Zeitouni et al., 2010
SVMerge	Pipeline integrating several existing callers followed by de novo assembly. Applied to the analysis of a HapMap trio	http://svmerge.sourceforge.net	Wong et al., 2010
SVSeq2	Split-read mapping for low-coverage sequence data	http://www.engr.uconn.edu/~jiz08001	Zhang et al., 2012
VariationHunter	Gives combinatorial formulations for the SV detection between a reference genome sequence and a NG-based, paired-end, whole genome shotgun-sequenced individual	http://compbio.cs.sfu.ca/strvar.htm	Hormozdiari et al., 2009

PEM, Paired-End Mapping; CNVs, Copy Number Variants; NGS, Next-Generation Sequencing; SVs, Structural Variants; SD, Segmental Duplication; WGS, Whole Genome Sequencing; pop., population; ind., individual; ref.,reference; seq., sequencing; ins, insertion; del, deletion; inv, inversion

seul chromosome chez *Miscanthus* ainsi qu'un certain nombre d'inversions (Ma et al., 2012). Chez le bananier, la très récente comparaison du génome B au génome A a révélé la présence d'une translocation réciproque impliquant deux fragments de respectivement 700 kb et 8 Mb sur les chromosomes 1 et 3, ainsi que la présence d'une inversion de 9 Mb sur le chromosome 5 (Jin et al., in prep.).

(iv) Séquençage de BAC ou d'extrêmités de BAC :

On peut identifier des variations structurales entre deux génotypes en séquençant l'extrémité de BAC chez un génotype et en les alignant sur le second génotype dont la séquence de référence est à disposition. Les BAC liant des régions chromosomiques distantes peuvent révéler la présence de réarrangements structuraux entre les deux génotypes. Ce type d'approche a été réalisé pour comparer la structure du génome du Gibbon à celle de l'homme (Carbone et al., 2006). Le séquençage complet des BAC chevauchant les zones de rupture de la synténie entre les deux espèces comparées renseigne sur la nature de l'instabilité génomique dans ces régions (Carbone et al., 2006; Girirajan et al., 2009).

(v) Re-séquençage :

La disponibilité de séquences de référence a entraîné un développement important des recherches sur les variations structurales, en particulier celles de petite ou moyenne taille, ce qui a entraîné le développement de différents types d'outils informatiques résumés dans le **Tableau 3**. Le principe de ces approches est de séquencer massivement une accession par NGS, d'aligner les séquences obtenues (lectures) sur le génome de référence et de chercher dans l'alignement de ces lectures des différences (discordances) par rapport à la séquence de référence.

6.3 La recherche de variations structurales par des approches de re-séquençage

Il existe actuellement trois principales approches pour détecter les variations structurales par re-séquençage (revues par Alkan et al., 2011a; Xi et al., 2010): les algorithmes basés sur la variation de couverture (nombre de lectures couvrant un site de la séquence de référence), les algorithmes basés sur la détection de lectures discordantes (lectures présentant des configurations d'alignement non attendues) et les algorithmes basés sur la détection de lectures découpées (lectures dont l'alignement n'est pas complet).

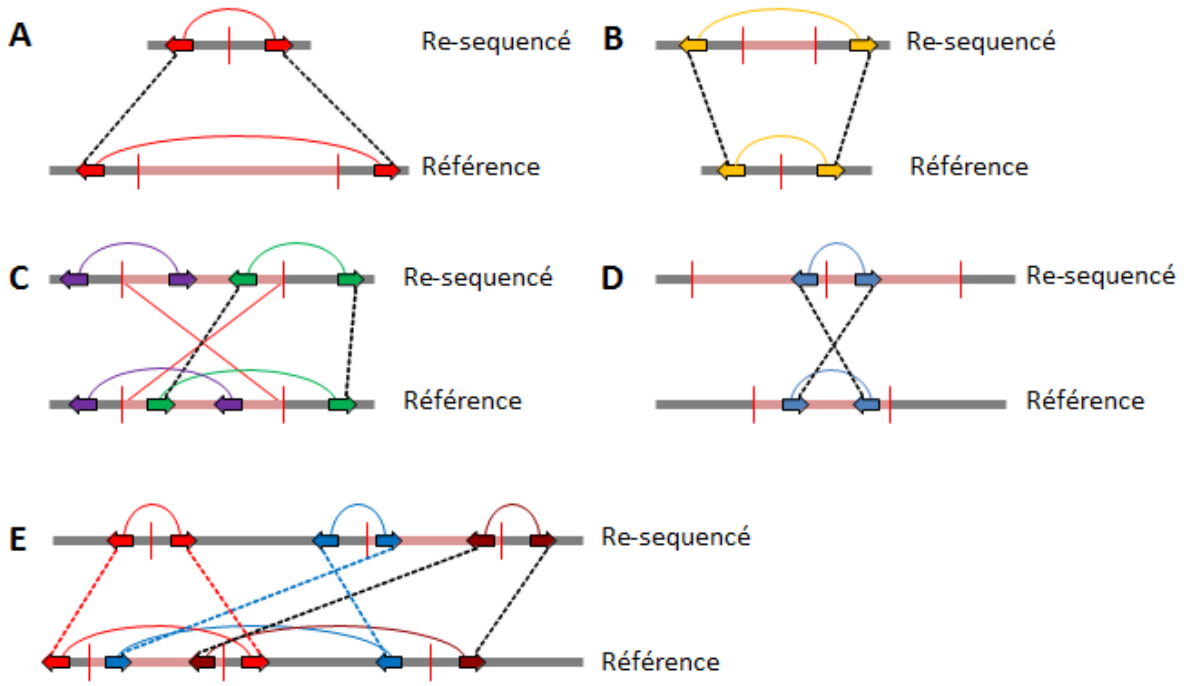


Figure 15 : Configurations des lectures pairées pour différents types de variations structurales. Les lectures sont obtenues par re-séquençage d'une accession puis alignées sur la référence. L'orientation attendue des paires est reverse (flèche vers la gauche) – forward (flèche vers la droite). (**A**) Délétion chez le re-séquencé : les lectures pairées (flèches rouges) chevauchent la zone manquante (trait rose) chez le re-séquencé mais présente chez la séquence de référence. La discordance se traduit par des lectures dont la taille d'insert est beaucoup plus importante qu'attendue. (**B**) Insertion chez le re-séquencé : les lectures pairées (flèches jaunes) chevauchent la zone qui a été « perdue chez la référence » (trait rose) et la discordance se traduit par des lectures dont la taille d'insert est plus petite qu'attendue. (**C**) Inversion chez le re-séquencé : les lectures pairées (flèches violettes et vertes) chevauchent la zone de rupture de la synténie liée à l'inversion (trait rose). La discordance se traduit par un changement de l'orientation d'une des lectures de la paire (reverse-forward vers forward-forward, flèches vertes ou reverse-reverse, flèches violettes) et peut être cumulée à une variation de la taille de l'insert par rapport à la taille attendue. (**D**) Duplication en tandem chez le re-séquencé : les lectures pairées (flèches bleues) à cheval sur deux motifs répétés (trait rose) chez le re-séquencé se retrouvent alignées sur la référence avec une orientation opposée (forward-reverse) à l'orientation attendue et peut être cumulée à une variation de la taille de l'insert. (**E**) Translocation : les lectures pairées (flèches de couleur) montrent alors plusieurs types de discordances. Les lectures chevauchant une extrémité de la zone transloquée (trait rose) chez le re-séquencé montrent une discordance de type délétion (bonne orientation mais trop grand insert, flèches marrons) alors que les lectures de l'autre extrémité (forward-reverse, flèches bleues) montrent une discordance de type duplication en tandem. Les lectures chevauchant la zone qui contenait le segment transloqué chez la référence se retrouvent avec une discordance de type délétion (bonne orientation mais trop grand insert, flèches rouges).

La première approche consiste à rechercher des régions génomiques où la couverture en séquences est significativement différente de la couverture attendue. Deux variantes peuvent être appliquées pour estimer la couverture attendue. Soit en utilisant la distribution moyenne des couvertures sur une fenêtre donnée (ex : EWT – Yoon et al., 2009), soit en utilisant un génome re-séquencé contrôle (ex : SegSeq – Chiang et al., 2009). Ces approches partent du postulat que l'échantillonnage des différentes zones d'un génome est complètement aléatoire. Or il a été montré qu'avec les technologies actuelles il y a des biais d'échantillonnage et d'alignement aléatoire liés :

- au pourcentage en GC (Dohm et al., 2008),
- à une meilleure accessibilité à certaines zones du génome, qui sont donc plus faciles à fragmenter et séquencer,
- ou encore une « mappability » (qualité de l'alignement) non constante au sein d'un génome (Derrien et al., 2012).

La « mappability » peut être estimée et prise en compte dans les algorithmes. De même, quelques algorithmes corrigent les biais liés à la composition en GC mais ceux liés à l'accessibilité du génome sont plus difficiles à estimer d'où l'intérêt de l'utilisation d'un génome contrôle.
Cette approche permet de détecter les indels ainsi que les variations du nombre de copies d'une région mais ne permet pas de détecter les translocations ou les inversions.

La seconde approche consiste à utiliser des lectures pairées (paires de séquences d'ADN obtenues par séquençage des extrémités d'un fragment d'ADN de taille connue). Pour la suite de cette thèse, le terme de "banque de x kb" désignera un ensemble de lectures pairées issues du séquençage des extrémités d'un grand nombre de séquences ADN de taille moyenne ("insert") de x kb. Les lectures sont ensuite alignées sur la référence et les lectures pairées dont la configuration d'alignement est différente de la configuration attendue en raison de variations structurales sont identifiées comme paires discordantes (**Figure 15**). Des configurations plus ou moins compliquées peuvent être trouvées en fonction de la complexité de la variation structurale mais aussi en fonction du rapport de la taille de la discordance sur la taille de "l'insert". Cette approche permet de détecter les insertions, les délétions, les inversions et les translocations inter- et intra-chromosomiques. Plusieurs programmes se basant sur cette approche existent parmi lesquels on peut citer SVdetect (Zeitouni et al., 2010), BreakDancer (Chen et al., 2009) ou encore VariationHunter (Hormozdiari et al., 2009).

La troisième approche, identifie des lectures qui présentent des problèmes d'alignement sur le génome de référence sur lequel elles sont alignées. Dans cette approche, la lecture contient la variation structurale et deux configurations peuvent être observées : i) la totalité de la lecture s'aligne mais en deux morceaux à plusieurs bases de distance, dans ce cas une délétion dans l'accession re-séquencée est détectée, ii) Les deux extrémités de la lecture s'alignent côte à côte mais pas sa région centrale, signifiant une insertion dans le génotype re-séquencé correspondant à la partie de la lecture qui ne s'aligne pas. Ce type d'approche fonctionne pour la détection d'insertions que si la lecture obtenue sur le génome re-séquencé est plus grande que la variation. Des variantes qui utilisent ces lectures s'alignant partiellement existent par exemple pour détecter des insertions plus grandes que la lecture. Ces variantes utilisent les lectures alignées partiellement pour identifier des zones de rupture et supposer une insertion. On peut par exemple citer Pindel (Ye et al., 2009) qui utilise en plus les lectures pairées pour réduire l'espace de recherche des lectures s'alignant partiellement.

7-Présentation du sujet de thèse

Les cultivars de bananiers sont dérivés d'hybridations naturelles entre espèces et sous-espèces sauvages de *Musa*. L'espèce *M. acuminata* est impliquée dans tous les cultivars et l'espèce *M. balbisiana* dans une partie d'entre eux.

L'analyse des figures d'appariements à la méiose suggère la présence d'hétérozygoties structurales chez les hybrides entre espèces et sous-espèces de bananier. La stérilité des bananiers cultivés est au moins partiellement attribuée à ces hétérogénéités de structure des chromosomes. D'autre part, les études de cartographie génétique réalisées jusque-là révèlent la présence d'importantes distorsions de ségrégations dont certaines semblent affecter des groupes de liaison/chromosomes entiers. L'origine de ces distorsions est souvent attribuée à ces variations de structure chromosomique sans que d'autres causes comme la sélection génique aient pu être exclues.

Les variations structurales, en affectant le fonctionnement normal de la méiose et la ségrégation des chromosomes, pourraient donc avoir un impact sur la fertilité des bananiers cultivés. D'autre part, elles pourraient être responsables des distorsions de ségrégation qui limitent les analyses du déterminisme génétique des caractères d'importance agronomique et qui affectent de façon importante les stratégies de brassage allélique dans les programmes d'amélioration. Si on veut préserver les possibilités de recombinaison récurrente pour cumuler des caractères, gérer l'hétérozygotie générale et utiliser de la stérilité hybride corrélée à

l'hétérozygotie structurale, il est très important de mieux appréhender les différences de structure chromosomique au sein des *M. acuminata*, leurs distributions par rapport à celle de la diversité génétique des bananiers et leur impact sur les ségrégations chromosomiques.

Dans ce contexte, la disponibilité d'une séquence de référence du génome de *Musa acuminata* (sous-espèce *malaccensis*, accession 'Pahang' haploïde doublé) constitue une opportunité importante pour tester de nouvelles approches pour caractériser les variations structurales chez le bananier.

L'objectif de la thèse était de mieux comprendre la nature des variations structurales au sein des *Musa acuminata* et leur impact sur la ségrégation des chromosomes en utilisant des stratégies de re-séquençage.
Ce travail a nécessité la mise en place d'approches et d'outils bioinformatiques pour exploiter les données de re-séquencage afin de détecter et caractériser les variations structurales de grande taille (translocations, inversions et duplications). Sur la base des premiers tests réalisés avec ces outils, il a été jugé nécessaire d'améliorer la séquence de référence de *Musa acuminata* (génome A) disponible. Cette amélioration a donc été réalisée et pour ce faire de nouvelles approches et de nouveaux outils bioinformatiques ont été développés.
Les approches de détections des grandes variations structurales mises en place ont ensuite été appliquées, avec la nouvelle version de la séquence de référence du génome A, à l'analyse de l'accession 'Pahang', le parent de l'haploïde doublé utilisé pour produire la séquence de référence du génome A et à l'accession 'PKW' qui a récemment été utilisée pour produire la séquence de référence de l'espèce *M. balbisiana* (génome B).
D'autre part, nous avons analysé en détails la ségrégation des chromosomes de l'accession 'Pahang' dans une descendance issue de son autofécondation et au moyen de marqueurs GBS de type DArTSeq. En effet, une première carte génétique de ce génotype avait montré la présence de fortes distorsions de ségrégation au niveau de deux chromosomes qui pouvait suggérer la présence d'une hétérozygotie structurale.

Le travail d'amélioration de la séquence de référence du génome A et des outils mis en place fait l'objet du chapitre I. Il est décrit sous la forme d'un projet d'article sur l'amélioration de la séquence de référence du génome nucléaire, un article paru dans la revue PlosOne sur l'assemblage et l'annotation du génome chloroplastique et un sous-chapitre sur l'identification des scaffolds correspondants au génome mitochondrial.

La description des approches et outils bioinformatiques mis en place pour détecter les variations structurales, leur application à l'accession 'Pahang' et à l'accession 'PKW' sont décrits dans le chapitre II.

L'analyse des distorsions de ségrégation dans l'accession 'Pahang' fait l'objet du chapitre III de la thèse qui est rédigé sous la forme d'un projet d'article.

Chapitre I : Amélioration de la séquence de référence du génome du bananier

I.1 Amélioration de la séquence de référence du génome du bananier

L'amélioration de la séquence de référence du génome du bananier a été réalisée en utilisant comme base les contigs et les données de séquençage générés à partir de l'haploïde doublé 'Pahang-HD' lors de la production de la séquence de référence du génome du bananier (D'Hont et al., 2012), auxquels ont été ajoutées de nouvelles données.
Ces données supplémentaires sont tout d'abord des lectures pairées obtenues à partir du séquençage d'une banques d'insert de 5 kb de 'Pahang-HD' correspondant à 21 Gb de séquence, soit une couverture additionnelle d'environ 40x ; puis des données de cartographie génétique obtenues par la méthodologie DArTseq de génotypage par séquençage (GBS) produites sur la population ayant servi à faire le premier ancrage des scaffolds sur la séquence de référence du génome du bananier (D'Hont et al., 2012) ; et enfin des données de carte physique du bananier 'Pahang-HD' obtenues en utilisant le système de cartographie optique Irys (http://www.bionanogenomics.com/technology/irys-technology/).

Un nouvel assemblage des contigs en scaffolds (scaffolding) a été réalisé avec l'outil SSPACE (Boetzer et al., 2011) à partir des contigs de séquences déjà existants et en ajoutant les nouvelles données de séquençage. Cette étape a permis d'obtenir un nouvel assemblage de même taille mais de meilleure qualité que le précédent.
J'ai ensuite exploité les données de cartographie génétique (anciennes et nouvelles) pour identifier des problèmes d'assemblage. Chaque scaffold contenant des marqueurs génétiques provenant de groupes de liaison différents, donc potentiellement de chromosomes différents, a été ciblé. Sur chacun de ces scaffolds (33), les lectures pairées ont été positionnées dans les zones où l'assemblage pose problème grâce aux outils développés et présentés dans l'article 1 de cette thèse. L'analyse visuelle de ces zones a permis d'identifier avec certitude les zones de rupture dans lesquelles la séquence avait mal été assemblée.

J'ai ensuite recherché, à l'aide de la visualisation des lectures des séquences pairées localisées sur l'assemblage initial, des évènements d'assemblage qui n'avaient pas été détectés lors de

l'étape de scaffolding à cause des paramètres issus d'un compromis entre qualité et quantité à l'échelle globale du génome. La majeure partie de ces évènements consistait en des séquences qui pouvaient être intégrées à l'intérieur d'une autre à la place d'une succession de bases inconnues (fusion de scaffolds) ou des séquences qui pouvaient être reliées bout à bout (jonction de scaffolds). Une fois ces évènements identifiés de façon automatique, les mêmes outils de visualisation que précédemment ont été utilisés pour valider manuellement ces événements. Une fois validés, les étapes de fusion et de jonction des scaffolds ont été réalisées.

Après cette phase, la taille des régions contenant des N dans les scaffolds a été ré-estimée dans l'assemblage. En effet, l'importante couverture du génome dont nous disposons avec les séquences pairées de la banque à 5 kb permet d'estimer avec précision la taille de la majeure partie des zones d'incertitudes dont la taille est inférieure à la taille de la banque.

Les données de cartographie optique ont ensuite été utilisées pour valider les jonctions et fusions réalisées et ces données ont aussi permis de grouper quelques scaffolds supplémentaires. Une étape de remplissage des régions contenant des N a ensuite été réalisée.

Parallèlement, 12 scaffolds correspondant au génome mitochondrial ont été identifiés (décrit en détail dans le point I-3), ils ont donc été retirés de l'assemblage nucléaire. Les scaffolds restants ont enfin été groupés et ordonnés en pseudo-molécules en utilisant une nouvelle approche d'ancrage sur les chromosomes. Contrairement à l'approche classique de cartographie génétique qui ordonne des marqueurs sur la base d'ordonnancements locaux de paires de marqueurs, la méthode que nous avons développée est basée sur l'organisation non pas de paires de marqueurs mais de blocs de marqueurs génétiques déjà positionnés sur les scaffolds. L'ordre des blocs les uns par rapport aux autres est déterminé par une méthode dérivée de l'approche d'UPGMA couplée à des tests de permutation. Cette méthode est implémentée dans deux programmes "*UPGMA"* et "*reorderient"* ces derniers ont été développés lors de cette thèse.

Pour finaliser la nouvelle version de l'assemblage, les annotations de gènes ont été transposées de la première version de l'assemblage à la nouvelle version de l'assemblage par Gaëtan Droc (CIRAD).

Grâce aux données supplémentaires que nous avons générées lors de ce travail et à leur analyse via les outils développés, la séquence de référence a été améliorée. Le nombre de scaffolds est passé de 7513 à 1520. La taille moyenne des scaffolds a été améliorée, en

passant de 63 kb à 294 kb, la N50 a plus que doublé passant à 3.0 Mb, le nombre de bases indéterminées contenues dans l'assemblage a diminué pour passer de 81.7 Mb (17%) à 45.2 Mb (10%) et la fraction ancrée du génome a été augmentée de 332 Mb (70%) à 397 Mb (89.5%).

La méthodologie appliquée pour améliorer la séquence de référence du bananier est décrite en détail sous la forme de l'article qui suit intitulé **A protocol to go beyond draft genome assemblies in plants: the banana sequence as a case study** (à soumettre prochainement).

Projet de publication n°1

Title

A protocol to go beyond draft genome assemblies in plants: the banana sequence as a case study

Guillaume Martin[1], Franc-Christophe Baurens[1], Gaetan Droc[1], Mathieu Rouard[2], Andrzej Kilian[3], Alex Hastie[4], Françoise Carreel[1], Angélique D'Hont[1*]

[1]. CIRAD (Centre de coopération Internationale en Recherche Agronomique pour le Développement), UMR AGAP, F-34398 Montpellier, France

[2]. Diversity Arrays Technology, Yarralumla, Australian Capital Territory 2600, Australia.

[3]. Bioversity International, Parc Scientifique Agropolis II, 34397 Montpellier Cedex 5, France.

[4]. BioNano Genomics, 9640 Towne Centre Drive, San Diego, CA 92121, USA.

[*]. Corresponding author: Angelique D'Hont

UMR AGAP, CIRAD, TA A-108/03, Avenue Agropolis, 34398 Montpellier cedex 5, France

Phone: +33 (0)4 67 61 59 27

Fax: +33 (0)4 67 61 56 05

Email address: angelique.d'hont@cirad.fr

Abstract

The recent advances in high throughput sequencing methods have led to a drastic increase in the number of released draft genome sequences. However, the completion of these newly sequenced draft genomes greatly varies and their quality remains an important challenge. Because assembly quality highly impacts subsequent exploitation of reference sequences, we implemented a semi-automated protocol to improve a reference sequence. We propose a modular approach that can use deep coverage of long insert-size paired-reads or genotyping by sequencing (GBS) of segregating populations, or Optical mapping data. Long insert size paired-reads are used to perform scaffold junctions and fusions left over by automatic assembly methods. GBS data can be used to detect and correct scaffold miss-assembly but also to anchor scaffolds into pseudo-molecules with a new approach that avoids the tedious step of marker ordering during genetic map construction. Finally, optical mapping data were used to assemble scaffolds into super scaffolds. Unlike automated methods proposed by assembler the semi-automated method described here proposes possible improvements that curators can validate through local compromises.

As a proof of concept, this approach has been used to produce a new version of the banana (*Musa acuminata*) reference genome sequence that was greatly improved. The total scaffold number was reduced of 79% (7513 vs 1532), with a N50 that increased from 1.3Mb (65 scaffolds) to 3.0 Mb (26 scaffolds). The unanchored sequences were reduced from 140.4 Mb to 46.5 Mb. These sequences were integrated in the 11 pseudo-molecules resulting in 89.5% of assembly anchored to the chromosomes compared to the previous 70%. Finally, unknown sites (N) were reduced from 17.3% to 10.0%.

Introduction

The two first and most complete plant genomes sequenced so far were *Arabidopsis* and Rice in 2000 and 2005 through the sequencing of a minimum tiling path of bacterial artificial chromosome (BAC) clones selected from a physical map. Since then the number of sequenced plant genomes has increased steadily each year thanks to the improvement of sequencing technologies in cost and speed (Bolger et al., 2014; Feuillet et al., 2011; Michael and Jackson, 2013). Nowadays, most genome assemblies are produced through whole genome shotgun sequencing (WGS) using Next Generation Sequencing (NGS) and even huge or complex genomes are now being sequenced (*e.g.* wheat: The International Wheat Genome Sequencing Consortium (IWGSC), 2014). However, one drawback has been that the quality of the assembly has generally decreased and is very variable among sequenced genomes.

The main steps that lead to a whole genome sequence assembly using the WGS approach are i) assembling of raw sequence reads based on their overlap into larger sequences called contigs; ii) building bridges between contigs using end-sequences from DNA fragments of various lengths (*e.g* BAC, fosmids, plasmids, large insert size libraries) to generate scaffolds; iii) anchoring scaffolds to chromosomes using genetic mapping data to produce the pseudo-molecules.

A major challenge toward long assembly is genome sequence redundancy that is particularly high in plant genomes. The main source of redundancy is repeated sequences in the form of transposable elements that represent a large part of plant genomes (from 14% in *Arabidopsis* to 80% in wheat) (reviewed in Kejnovsky et al., 2012). Another source is the abundance of paralogous genes (Hahn et al., 2014) resulting from various types of duplications including whole genome duplications that occurred frequently in plant genomes (Vanneste et al., 2014). During the assembly steps, repeated sequences are often assembled into single collapsed region (Alkan et al., 2011) leading to conflicts in sequence assembly with two problematic options for automatic assemblers: miss-assembly of non contiguous regions or premature stop of sequence assembly process.

These constraints are exacerbated with short insert size reads since the insert size most of the time does not allow to span the repeats. Conversely, scaffolding with only very large insert size libraries (i.e. BAC-end libraries) limits the integration of small scaffolds in the final assembly.

Assembly quality highly impacts subsequent exploitation of reference sequences. Remaining errors in the assembly, scaffolds ordering or anchoring, can lead to misinterpretation when studying genome structure. They also decrease the efficiency of gene predictors (Alkan et al.,

2011), and impact genome wide association and population genetic studies (Mascher and Stein, 2014). In addition, high proportion of sequences not anchored to chromosomes usually referred to as chr0 or chr_unknown also decreases the power of comparative genomics analysis. Chain et al., (2009) defined a set of standards allowing evaluation of the quality of assembled genome from *standard draft* to *finished* whole genome sequences and identified subsequent analysis that can be performed depending on the completion of the genome.

New approaches are continuously being developed that can be used to improve genome reference assemblies. Among the recent ones are high coverage medium and large insert size library (Mardis, 2011; Williams et al., 2012), genotyping by sequencing for assembling scaffolds into pseudo-molecules (Mascher and Stein, 2014; Mascher et al., 2013) and construction of physical maps from single DNA molecule (optical mapping (Dong et al., 2013; Neely et al., 2011), Irys genome mapping (http://www.bionanogenomics.com/technology/irys-technology/, Levy-Sakin and Ebenstein, 2013). However, in contrast to the tremendous advances in sequencing throughput, assembling sequences remains a substantial endeavor (Schatz et al., 2012). Several automated programs have been developed to improve draft genome assemblies such as Bambus (Pop et al., 2004), SOPRA (Dayarian et al., 2010), MIP (Salmela et al., 2011), SSPACE (Boetzer et al., 2011), Opera (Gao et al., 2011), GRASS (Gritsenko et al., 2012), SCARPA (Donmez and Brudno, 2013), SSPACE-LongRead (Boetzer and Pirovano, 2014), SOAP-de-novo2 (Luo et al., 2012), GapFiller (Boetzer and Pirovano, 2012) and PAGIT (Swain et al., 2012). In this paper, we describe a set of new semi-automated tools that can be used in combination with some of these already available automated programs for improving draft genome sequences at the different steps of the assembly process. Unlike automated programs in which parameter compromises are set up at the genome scale, the semi-automated method presented here points out possible improvements to the user who can validate them through local compromises.

The described method was applied to improve the current banana (*Musa acuminata*) whole genome sequence (D'Hont et al., 2012) at various levels: i) correct miss-assemblies and increase scaffold N50, (2) decrease the proportion of unknown sequences (N) iii) reduce the proportion of unanchored scaffolds. The method can also be applied during *de novo* assembly and can be used as a tool to validate quality of a final assembly.

To ensure that the developed tools are accessible to a wide community of users, we made them available on our Galaxy platform (http://gohelle.cirad.fr/galaxy//) under *Scaffhunter* and *Scaffremodler* tools and a command line version is available on GitHub

Table 1: Data requirrement to perform the different whole genome assembly improvment step

			Data requirement								
	Steps*	**improvement steps name**	contigs	scaffolds	pseudomolecule	small insert size libraries (< 1kb)	large insert size libraries (> 1kb)	Genotyping by sequencing	Physical map	annotation
scaffold construction	step1	(Re-)scaffolding of contigs	**X**				**X**			
Scaffold improvement	step2	Identification and splitting of scaffolds/contigs miss-assembly		**X**			**X**	**X**[b]	**X**[b]	
	step3	Scaffold fusion/junctions		**X**			**X**		**X**[b]	
	step4	Scaffold gap re-estimation		**X**			**X**			
	step5	Super scaffold construction		**X**					**X**	
	step6	Scaffold gap closure		**X**		**X**				
Assembly in pseudomolecule	step7	Scaffold anchoring		**X**				**X**		
	step8	Annotation transposition		**X**[a]	**X**[a]					**X**

* : steps are named in case of availability of all datasets required

[a] : either one or the other data is required

[b] : At least one data type is required but they can be used together

https://github.com/SouthGreenPlatform. Galaxy is among the leading graphical systems for executing workflows and visualizing sequencing assay results (Schatz et al., 2012).

Material and methods

Methods

Table 1 provides an overview of the method for improving draft genome sequences. It is divided into distinct steps that can be performed independently depending on available data and objectives.

Step1 consists in performing a new scaffolding of existing contigs by adding data from high coverage paired sequences from large DNA segments (HCPSL). Step2 consists in identifying and splitting miss-assembled contigs/scaffolds using a combination of GBS genetic mapping data and HCPSL. Step3 uses HCPSL to perform scaffold fusions and junctions. Step4 uses all paired sequences (*e.g*. HCPSL, BAC-ends sequences and plasmid paired reads) to re-estimate gap size. Step5 consists in an additional scaffold grouping step using physical mapping data from a Irys genome map. Step6 attempts to close gap regions using illumina paired reads of small insert size libraries. Step7 consists in ordering and assembling scaffolds into pseudo-molecules using GBS data from a segregating population. Step8 consists in transposing annotation from the first draft genome version to the new assembly. The programs we developed for performing steps 2, 3, 4 and 7 are available on the Galaxy platform under Scaffhunter and Scaffremodler tools.

Step1: Re-scaffolding of contigs

Contigs are assembled into scaffolds using *SSPACE* (Boetzer et al., 2011). The scaffolding process is divided in as many steps as the number of sequenced libraries with distinct inserts sizes. The libraries are used by increasing insert size order, scaffolding parameters are optimized for each step and use the scaffolds obtained from the previous one. To prevent accumulation of scaffolding errors, the first library is used with more stringent parameters (-a 0.5, -k 20) than the second and third one (-a 0.7, -k 1). For Sanger sequence libraries (*e.g.* BAC-end sequences, fosmids) reads are mapped as single end reads using *BWA* (Li and Durbin, 2010). Read pairs are then reconstructed and stored in a tabulated file used by *SSPACE*.

Step2: Identification and splitting of scaffold/contig miss-assemblies

This step aims at identifying and splitting miss-assembled scaffolds using GBS data from segregating populations and paired reads obtained from HCPSL. The complete process and tools used are summarized in Supplemental Figure S1.

Step2 can be decomposed as follows: markers are attributed to linkage groups using JoinMap4.1 software (van Ooijen, 2011) and standard grouping parameters; In parallel, marker sequences are aligned to scaffolds using the *locOnRef* tool; At the end of these two steps, scaffolds harboring markers attributed to more than one linkage group are further investigated for evidence of miss-assembly. To search for evidence of miss-assembly, mate-pair reads located in scaffolds comprising genetic markers from different linkage groups are plotted to generate visual data of sequencing coverage at a local scale. The principle is to identify all reads located on these scaffold regions, draw pairing information by joining both reads of a pair and detect a region with no pair overlap that may indicate miss-assembly.

Paired reads are first mapped onto scaffold regions using *1_create_conf* and *2_map* tools. Identical read pairs (duplicates) and reads having multiple hits are filtered out with the *3_filter_single_pair* tool and *4_filter_sam* tool, respectively. Statistics such as median insert size that is used to re-estimate correctly mapped reads are calculated at this step and scaffold coverage and proportion of discordant reads are also calculated with the *5_calc_stat* tool. Read pairs are then parsed according to their orientation and insert size with the *6_parse_discord* tool. Finally, configuration files with data on reads pair links, coverage and discordant proportion in scaffold regions are generated with *conf4circos* tool and circos pictures (Krzywinski et al., 2009) are generated using the *draw_circos* tool.

Zones that are not spanned by read-pairs are visually identified and coordinates of sequence breakpoints are identified using a coverage file generated by the *5_calc_stat* tool. Once all scaffold splitting zones are identified, scaffolds are split using *convert2X* and *SplitOnX* tools. The final file is a multi-fasta file containing all scaffolds, including the newly split scaffolds. All scaffolds are renamed by decreasing order of length for step3 processing.

Step3: Scaffold fusions/junctions

This step aims at identifying scaffolds that may be contained into larger ones especially in zones containing stretches of N (hereafter referred to as fusion) and scaffolds that should be end-joined (hereafter referred to as junction). First, potential scaffold fusions and junctions are identified using discordant reads and then scaffold fusions and junctions are independently

and sequentially treated. The complete process and tools used are summarized in (Supplemental Figure S2).

To identify scaffold fusions and junctions, reads are mapped onto all scaffolds obtained from step2 and filtered out from redundancy. Mapping parameters are adjusted with local statistics (*i.e.* median insert size) as in step2. Reads are then parsed according to their orientation and insert size with the *6_parse_discord* tool. Discordant zones (*i.e.* zones that include discordant reads in wrong orientation or with incorrect insert sizes) are identified with the *7_select_on_cov* tool. Configuration files with data on discordant reads pairs, discordant proportion in scaffold regions and coverage are generated with *conf4circos* tool (already described in step2). Tab files containing putative fusion and junction zones and corresponding circos pictures are generated using *look4fusion* tool.

Candidate scaffold fusion zones are manually validated by ensuring that read pairs linking scaffolds are correctly orientated on each circos pictures. Scaffolds are then merged (scaffold fusion) using *fusion_scaff* tool. Once scaffold fusions are performed, fusion zones are verified by running the pipeline a second time from *1_create_conf* to *6_parse_discord* tools on the newly merged scaffolds. Circos configuration files are created using *conf4circos* tool and circos figures representing paired read link at each scaffold fusion boundaries are drawn with *verif_fusion* tool. Each fusion can be then visually validated by observing correctly orientated read pairs, overlapping the newly assembled fusion zone.

Candidate scaffold junction zones are manually validated by ensuring that read pair linking scaffolds are correctly orientated on each circos picture. To manage multiple scaffold joining, scaffolds are first grouped using *group4contig* tool that creates a table summarizing pairwise links and respective orientations of scaffolds within groups. This file is manually re-formatted to create a single scaffold order for each group based on pairwise data. This step cannot be automated because, in several cases, some links are missing (especially for small scaffolds and those containing repetitive sequences) and this can lead to ordering errors if automated. Scaffolds are then joined using *contig_scaff* tool.

Scaffold junctions are checked by running again the pipeline from *1_create_conf* to *6_parse_discord* tools on newly joined scaffolds. Configuration files for circos are created using *conf4circos* tool and circos pictures representing paired read links at scaffold junction boundaries are drawn with *verif_fusion* tool. Resulting circos figures are inspected for scaffold junction verification by observing correctly orientated read pair overlapping the newly assembled junction zone.

If a scaffold is the object of single operations (junction or fusion), all these steps can be performed. If multiple events affect the same scaffold (e.g. two scaffolds should be joined and further integrated into a third one, cascade of scaffold fusions), steps should be performed sequentially by running again the whole pipeline. As an example, if multiple small scaffolds have to be integrated into one same region, these small scaffolds should be either grouped together first (using *contig_scaff* tool) and then integrated as a single scaffold, or sequential fusion steps should be performed using the complete process from *1_creat_conf* to *fusion_scaff* tool until there is no fusion left.

The verification involved mapping and read filtering that are time consuming steps. To save time, these verifications steps have been performed on randomly reduced set of paired-reads representing ¼ of the complete dataset.

Step4: Scaffold gap re-estimation

As the scaffold junction and fusion step does not re-estimate gap size, the size of all remaining scaffold regions including N (gaps) is re-estimated using HCPSL data. The complete process and tools used are summarized in Supplemental Figure S3.

For this step, paired reads are mapped to the scaffolds using *1_create_conf* and *2_map* tools. Reads having multiple hits and read pair duplicates are filtered out with the *3_filter_single_pair* tool and *4_filter_sam* tool respectively. Gap sizes (number of N) are re-estimated with the *reEstimateN* tool. This tool uses correctly orientated read pairs that overlap gap regions to re-estimate the size of these gaps. To use this workflow several times with multiple libraries and to prevent re-estimating already estimated gap regions, the pipeline generates re-estimated gaps with an "E" character for undetermined bases (instead of a classical "N" character). At the end of a gap size re-estimation process, "E" are replaced by "N" with a simple unix command line in the resulting multi-fasta file.

Step5: Super scaffold construction

This step exploits physical mapping data from an Irys (http://www.bionanogenomics.com/technology/irys-technology/) genome map (physical maps from single DNA molecule) to group scaffolds into super scaffolds. In this step, grouped scaffolds were separated by Ns corresponding to their expected distance in the physical map. This type of data could be exploited earlier in the process, for example in step2 in addition to the genetic markers to identify miss-assemblies in scaffolds or in step3 to perform scaffold junctions.

Step6: Scaffold gap closure

This step aims at filling gaps generated during scaffolding, using paired reads from short insert size libraries. This step is performed with GapCloser v1.12 program (Luo et al., 2012). At the end of this step, all scaffolds are renamed according to their length.

Step7: Scaffold anchoring

This step aims at grouping, ordering and orientating scaffolds into pseudo-molecules using GBS markers from a segregating population. GBS is often associated with a residual level of genotyping errors, that may lead to markers miss ordering within high density clusters of markers. Methods to circumvent this problem by correcting the dataset have been proposed (Spindel et al., 2013) and we choose to avoid the step of genetic map construction and a subsequent conciliation between genetic map and scaffolds in our approach. The principle here is to find the best order within each linkage group (LG) using blocks of markers. These blocks of markers, that correspond to scaffolds, are first ordered relative to each other, using UPGMA-like based approach, and this first order is improved with permutation testing. The complete process and tools used are summarized in Supplemental Figure S4. Within each linkage group, all pairwise linkage LODs between markers are calculated using JoinMap4.1. The original pairwise file is converted into pairwise matrix with *JMpwd2matrix* tool. In parallel, sequence of markers is mapped onto scaffolds using the *locOnRef* tool. A first order is calculated using an UPGMA like approach on mean pairwise linkage LOD calculated between scaffolds with the *UPGMA* tool. Final scaffold ordering and orientation are optimized by performing scaffold permutations and re-orientations leading to maximization of a score calculated as follows:

$$\text{score} = \sum_{i=1, j=1,\, x_i<x_j}^{n} \left(1 - \frac{(x_j - x_i)}{\text{n}-1}\right) \text{LOD}_{\text{ij}}$$

with n the number of markers in the LG to order, x_i and x_j are the position of markers i and j in the tested order, and LOD_{ij} the LOD score between markers i and j. To optimize computation time and as order is not tested within scaffolds, i and j are markers from different scaffolds.

The optimization is performed with the *reorderient* tool. Scaffold sequences are then assembled into pseudo-molecule using the *scaff2chrom* tool. In addition to a fasta file containing ordered scaffold sequences separated by 100 N, an AGP file locating scaffolds into

pseudo-molecules is also generated. A dot-plot showing marker linkages along pseudo-molecules is performed using *matrix2ortho* tool and inspected for scaffold miss-ordering.

Step8: Annotation transposition

Gene annotations are transferred to the new assembly using Exonerate software (Slater and Birney, 2005). Exonerate performs genomic searches and spliced alignments in a single run. A custom Perl script takes this output and the previously GFF3 files to transfer the annotation on a new GFF3 files, and gives also a tabulated file with gene information linking the two releases.

Materials

Sequence data

454 contigs, Sanger 10 kb paired-reads and Sanger BAC-ends used for the production of the first version of the banana (*Musa acuminata*) reference genome and 330 b pair-end illumina sequences used to correct errors in this first reference sequence (D'Hont et al., 2012) were used. This reference sequence was produced with a doubled haploid of the 'Pahang' accession (DH-Pahang). An additional 40x 5 kb mate-pair library of DH-Pahang was generated and sequenced using the illumina HiSeq 2000 sequencing system at Genoscope (France). Reads were trimmed and filtered following three criteria: (1) trimming of both read ends until base quality is higher or equal to 20; (2) read trimming at the second unknown base in the sequence; and (3) read larger or equal to 30 bases were conserved.

Physical mapping data

A physical map of DH-Pahang accession has been constructed using the Irys system (http://www.bionanogenomics.com/technology/irys-technology/). The *de novo* map assembly yielded 464 Mb with a N50 of 715 kb (to be completed with Alex Hastie).

Genetic mapping data

A total of 180 individuals from a population of 268 individuals (AF-Pahang) obtained at the CIRAD research station in Guadeloupe from the self-progeny of the 'Pahang' accession (ITC0609) were genotyped using the DArTseq technology (Cruz, 2013). A total of 9,968 co-dominant (SNP) and 16,233 dominant markers were generated using a *Pst*I-*Mse*I enzyme combination. These markers were used in addition to the 768 SSR and 497 DArT markers previously used to anchor the *Musa acuminata* genome assembly. The 268 individuals of the

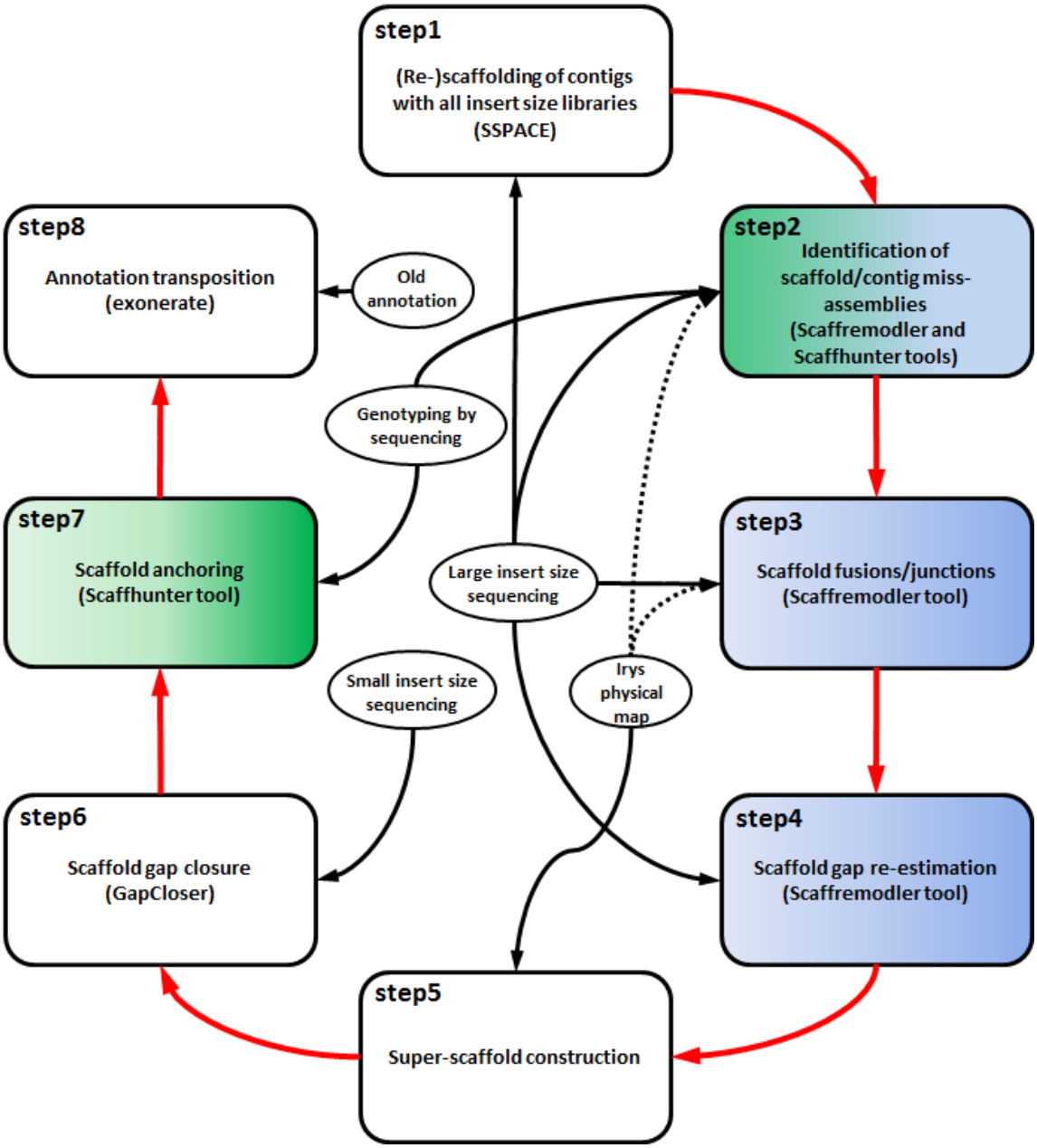


Figure 1: Methodology overview for the improvement of the *Musa* whole genome sequence. This method starts with contigs from a previous assembly. Required data are represented in the center of the figure in bubbles (Large insert size sequencing: mate-pair sequences from large DNA insert of different size and coverages, Genotyping by sequencing: markers generated on a segregating population, Small insert size sequencing: Illumina® 500bp paired reads, Irys® physical map: optical mapping based physical map).
Boxes colored in blue and/or green use programs available in *Scaffremodler* and *Scaffhunte*r tools respectively.
This methodology is divided in seven steps. In **step1,** contigs generated from the previous assembly are assembled into scaffolds with *SSPACE* using all mate-pair/pair-end libraries available. In **step2**, miss-assembled scaffolds/contigs are identified using a combination of paired reads and genotyping by sequencing (GBS) markers from a segregating population. In **step3,** identified miss-assembled scaffolds are split and scaffold fusions and junctions are performed using a high coverage large insert size library. In **step4**, using the high coverage large insert size library, the size of gaps are re-estimated. In **step5**, an additional step of scaffold merging is performed using a physical map constructed from the Irys® system. These data could also be integrated in step2. In **step6**, unknown sites are filled using *GapCloser* leading to the final scaffolds. The final scaffolds are then ordered into pseudomolecules using the GBS data in **step7**. In a final **step8**, annotation is transferred from the first to the second assembly using *exonerate*.
Red arrows indicate order for which different improvement steps were performed for the *Musa* genome sequence. Black arrows indicate datasets used for improvement and dotted arrows indicate where datasets could have been used for improvement.

mapping population were not genotyped with all types of markers: 91 individuals were genotyped with all types of markers, 178 individuals were genotyped with both DArT and DArTseq markers, 91 individuals with both DArTseq and SSR markers and 176 individuals were genotyped with both DArT and SSR markers. Markers were filtered, independently for each marker type, on the basis of the following criteria: no more than 20% missing data, no less than 10% heterozygous or dominant and no less than 1.5% homozygous for at least one homozygous state, resulting in 23,430 markers. The choice of these relatively non-stringent parameters was motivated by large segregation distortions that were previously observed in chromosome 1 and chromosome 4 in the segregating population (D'Hont et al., 2012). Only markers that could be uniquely mapped on the scaffolds were kept.

Results and Discussion

The method described above has been applied to improve the banana *Musa acuminata* reference genome assembly (D'Hont et al., 2012) using tools and datasets as summarized in Figure 1.

Musa contig scaffolding

The 24,425 contigs published in the original version of the *Musa acuminata* reference genome (D'Hont et al., 2012) were re-assembled into scaffolds using SSPACE (Boetzer et al., 2011) as described in step 1. Three distinct insert-size libraries were used: a 10 kb insert size Sanger library (3.4x coverage) and a BAC-end Sanger library (0.2x coverage), both used for the original version of the assembly, and a new 5 kb mate-pair illumina library (40x coverage).

Contigs were assembled into 2,267 scaffolds for a cumulated size of 439 Mb representing 84% of the estimated size (523 Mb) of the DH-Pahang genome. Ninety percent of the assembly was in 416 scaffolds and the N50 was 1.55 Mb. Gaps in scaffold represent 48.3 Mb accounting for 11% of the assembly. The important reduction of scaffold number between the first version of the assembly (7,513) and this version highlights the importance of median insert size library during the scaffolding step.

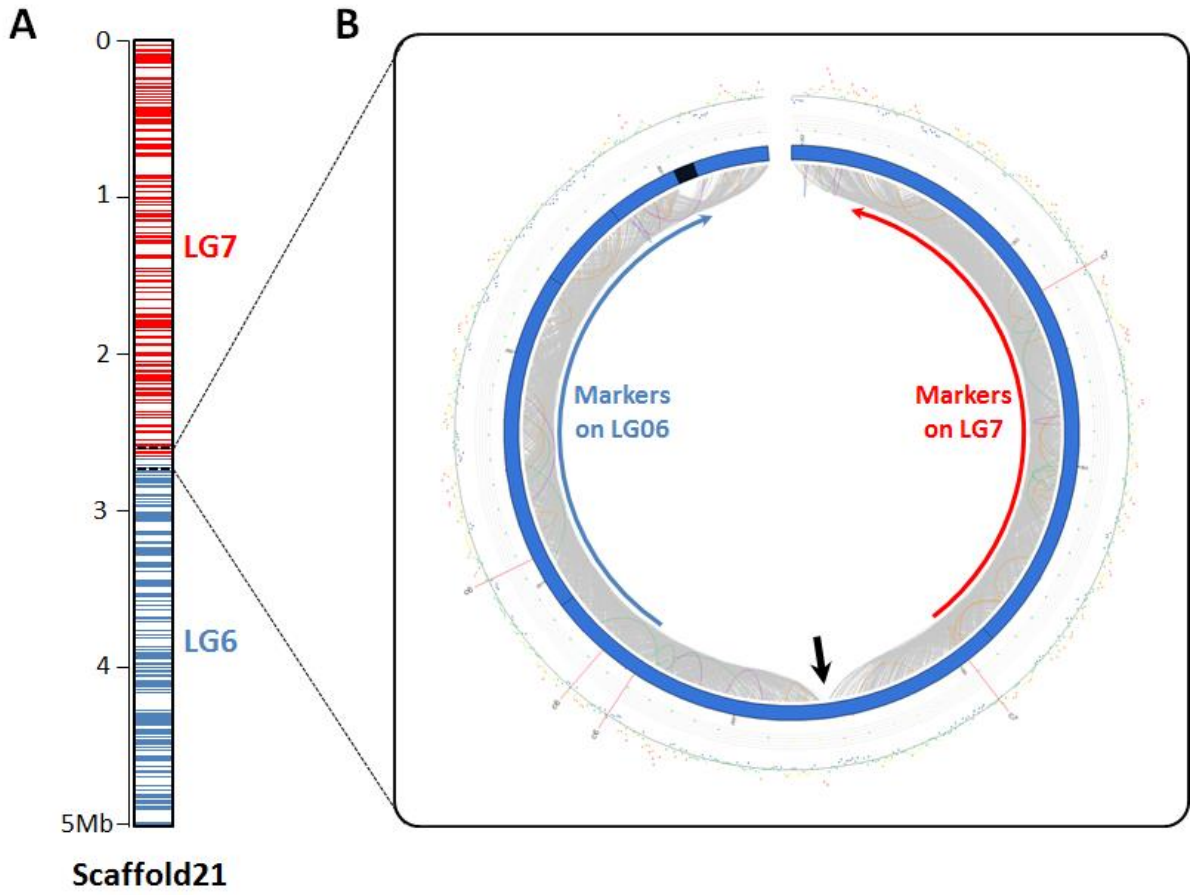


Figure 2: Example of clue leading to scaffold splitting. (**A**) Mapping of genetic markers onto scaffold21 of *Musa acuminata* after **step 1** (scaffolding). Markers in red and blue belong respectively to linkage-group 7 (LG7) and linkage-group 6 (LG6). (**B**) Circos graphical representation of paired read mapping in the region of scaffold21 where linkage-group shift is observed (dashed lines). This representation is drawn using *Scaffremodler* tool. In the inner circle, links between read pairs are drawn with the following color code: grey lines correspond to concordant pairs (correct orientation and insert size), orange and red lines correspond to discordant pairs with smaller and greater insert size respectively. Purple lines correspond to pairs showing reverse-reverse orientation, green lines, forward-forward and blue lines correspond to pair with complete reverse orientation relative to the paired library construction. The second circle represents scaffold in blue with gaps as black regions. The next circles are scatter plots with warm-cold color code. The first scatter plot presents the proportion of discordant reads on window size of one third of expected read pair insert size. The outer circle represents a scatter plot of read coverage on window size of 100 bases. Marker sequences aligned to this region of scaffold21 are noted c7 and c6 for LG7 and LG6 respectively. Scale is expressed in 10 kb unit numbered from the beginning of the scaffold. The black arrow identify the miss-assembled region in scaffold21 leading to the assembly of two regions that are not linked. Red and blue arrows indicate the two portions of the scaffolds belonging to LG7 and LG6 respectively.

Musa scaffold correction

The new scaffold assembly was subjected to additional corrections, fusions, junctions using methods and tools described in *Step2, 3, 4*.

- Identification of miss-assembled scaffolds (step2):

This step uses the genetic mapping data and the 5 kb mate-pair illumina library. A total of 21,942 markers uniquely mapped to the scaffolds were grouped into 11 linkage groups using JoinMap 4.1. After identification of scaffolds containing markers from distinct linkage groups 36 miss-assembled regions involving 33 scaffolds were identified. The miss-assembled regions were confirmed with the mapping of the 5 kb mate-pair illumina library on the assembly. The 36 miss-assembled regions were split resulting in a total 2,303 scaffolds. Figure 2 shows an example of a miss-assembled scaffold. The zone along the scaffold where marker linkage group shifts, indicates the potential miss-assembly (Figure 2A). Read pair mapping on the miss-assembled candidate region is then drawn using *Scaffremodler* tool. The breakpoint is identified by the absence of read-pair overlap of the area (Figure 2B) and an increased proportion of discordant reads.

Most of the zones identified here (24/36) can be assigned to scaffolding problems, potentially resulting from BES or Paired-end plasmid sequences that are probably issued from chimeric clones. The remaining zones (12/36) split sequences within contigs potentially indicating wrong assembly with the original 454 sequences. All these zones account for only a very small proportion (0.15%) of the original contigs, highlighting the quality of the primary assembly.

- Scaffold fusions and junctions (step3):

The scaffolding step always represents a compromise between parameters adapted to whole genome and local evidences. Global parameters are chosen to reduce false assembly but may locally prevent scaffold extension. In this step, discordant paired-reads from the 5 kb HCPSL were analyzed to identify possible scaffold fusions and junctions. A total of 438 scaffold fusions and 293 scaffold junctions were performed and validated, resulting in a reduction of scaffold number from 2,303 to 1,572. Figure 3 shows an example of scaffold fusion identified by *Scaffremodler* tool at step3. Figure 3A is an intermediate output of *step3* (automatically performed by *Scaffremodler* tool) showing read pair linking scaffold1112 and scaffold24. Automatic interpretation of the orientation of these links by *Scaffremodler* leads to the identification of a probable fusion of scaffold1112 into scaffold24 that should be validated by

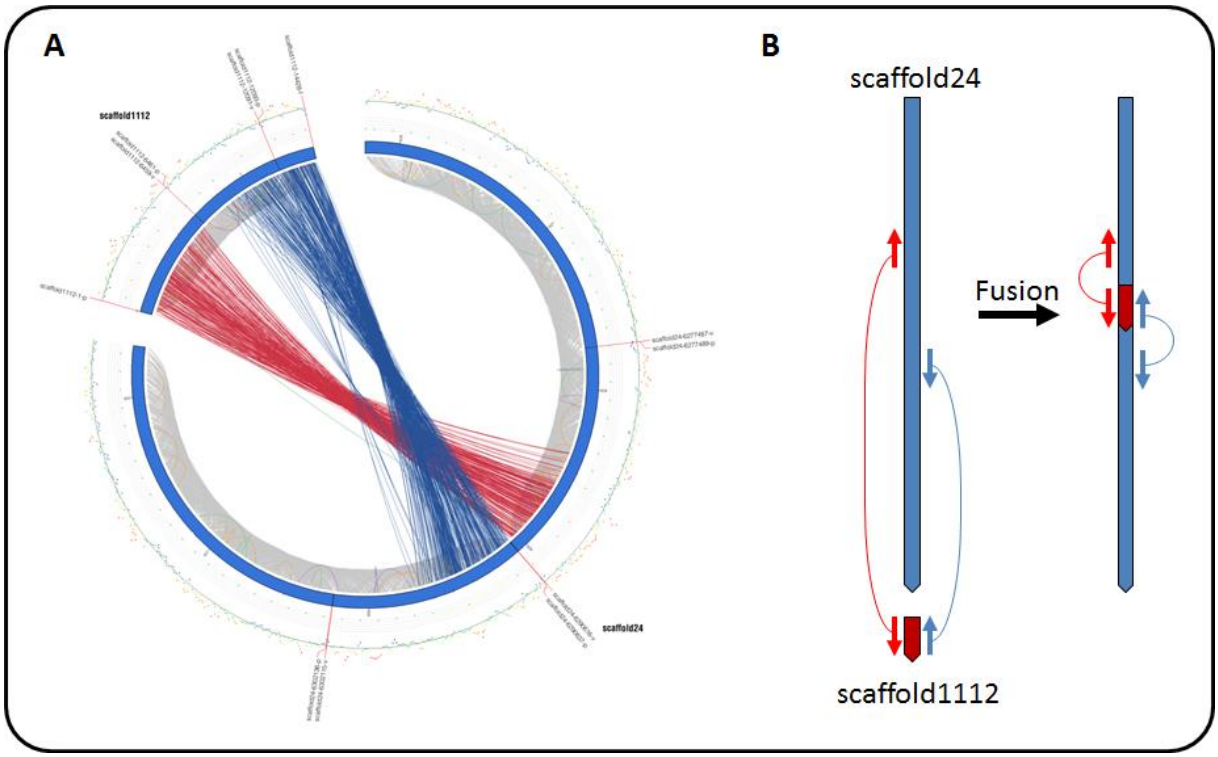


Figure 3: Example of clue leading to scaffold fusion. (**A**) Graphical representation of paired read leading to the identification of fusion of scaffold1112 into scaffold24. This representation is drawn using *Scaffremodler* tools. In the inner circle, links between read pairs are drawn with the color code described in Figure 2: grey for concordant pairs; red and orange for discordant in size; purple, green and blue for orientation discordant pairs. The second circle represents scaffold in blue with gaps as black regions. The next represents the proportion of discordant reads and the last circle represents read coverage as in Figure 2. Text locates start-1 (v) and end+1 (p) of gap regions in scaffold. Scale is expressed in 10 kb unit numbered from the beginning of the scaffold. Red and blue beams linking scaffold1112 and scaffold24 allowed identifying scaffold fusion schematized in (**B**). Inserting scaffold1112 into scaffold24 will correct the discordant red links and correct the orientation of discordant blue links.

Table 2: Statistics on scaffolds assembly

	V1 (D'hont et al 2012)	SSPACE	Fusion / joining / splitting / gap re-estimation	IRIS scaffold	GapCloser
scaffold number	7,513	2,267	1,572	1,532	1,532
cumulated size	472,210,317	438,736,528	443,852,100	450,994,104	450,697,673
unknown sites (%)	81,728,542 (17.3)	48,267,272 (11.0)	53,378,493 (12.3)	60,520,497 (13.4)	45,175,659 (10.0)
N50 (scaffold number)	1,311,088 (65)	1,545,585 (52)	2,890,075 (28)	3,014,384 (26)	3,016,874 (26)
N80 (scaffold number)	316,579 (299)	370,770 (242)	491,628 (169)	578,880 (150)	579,793 (150)
N90 (scaffold number)	54,335 (647)	169,980 (416)	201,127 (305)	234,686 (268)	234,825 (267)

the user. The specific orientation of scaffold1112 into scaffold24 solves the discordant links by correcting orientation of blue links and insert size of red links in the resulting sequence (Figure 3B).

After each step of fusion or junction, resulting scaffolds have been verified using mapped paired-reads as described in supplemental Figure S2, to ensure that the newly created junctions are spanned with reads mapped in the correct orientation. Supplemental Figure S5 shows the mapping of reads on the two borders of scaffold1112 after fusion into scaffold24. Both right and left borders displayed overlapping reads in the correct orientation (Supplemental Figure S5, A and B).

-Scaffold gap re-estimation (step 4):

The size of gaps within the 1,572 scaffolds was re-estimated using the paired reads libraries sequentially. First, the 5 kb mate-pair illumina library was used with a minimum of 30 concordant reads needed to re-estimate a gap, second, the 10 kb Sanger library (with at least 2 reads well orientated needed to re-estimate a gap), and third, the Sanger BAC-end library (with at least 1 read well orientated needed to re-estimate a gap). The cumulated size of the 1,572 scaffolds after gap re-estimation was of 444 Mb. Ninety percent of the assembly was in 305 scaffolds and the N50 was 2.9 Mb. Gaps in scaffolds represented 53 Mb for 12.3% of the assembly (Table 2).

Musa super-scaffolds (step5)

A physical map of DH-Pahang accession has been recently constructed using the Irys system (to be completed with Alex Hastie) and was exploited to group scaffolds into super scaffolds. The scaffold assembly was aligned to these genome maps and allowed merging of 72 scaffolds into 40 super-scaffolds, reducing scaffold number from 1,572 to 1,532. This physical map also validated scaffold junctions performed when possible (*i.e.* joined scaffolds are sufficiently large to be detected by the physical map). A total of 7.1Mb gap regions were added during super scaffold construction. Due to recent production of this map, we have used it only at this step but we anticipate that it could be used in addition to the genetic markers in step2 to identify miss-assemblies in scaffolds. Finally, 90% of the assembly was in 268 scaffolds and the N50 was 3.0 Mb with 26 scaffolds. Gaps in scaffolds represented 60.5 Mb for 13.4% of the assembly.

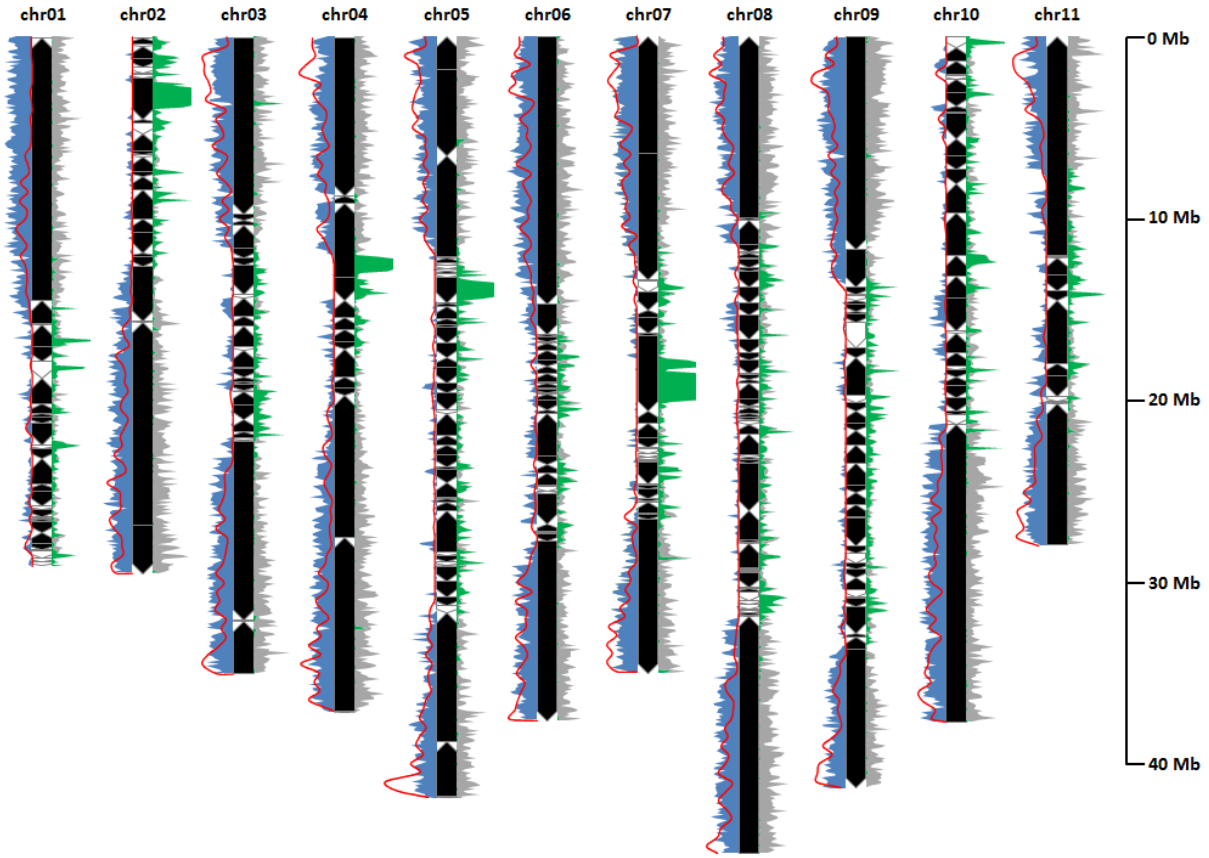


Figure 4: Density of markers, unknown sequences and genes and recombination number along the new version of eleven pseudo-molecules of *Musa acuminata*. The marker density (grey area) has been calculated as the number of grouped and mapped markers on windows size of 100 kb along each pseudo-molecule. Gene and unknown sequence density (blue area and green area respectively) have been calculated as the number of bases covered by genes and unknown sequences respectively, on windows of 100 kb. The recombination number (red curve) has been calculated as the number of recombinations observed in the population of 180 individuals within a sliding window of 500 kb on corrected SNP markers. Boxes symbolize scaffolds and their orientation along the pseudo-molecules.

Table 3 : Statistics on marker density on linkage groups

linkage group	cumulated scaffold size	nb_mark	mark_nb/100kb
chr01	29,067,552	1,384	4.76
chr02	29,509,134	1,502	5.09
chr03	35,017,413	1,920	5.48
chr04	37,104,143	2,489	6.71
chr05	41,848,132	1,924	4.60
chr06	37,589,864	2,234	5.94
chr07	35,025,021	1,744	4.98
chr08	44,883,571	2,728	6.08
chr09	41,302,925	2,136	5.17
chr10	37,671,811	2,023	5.37
chr11	27,952,850	1,519	5.43
total	396,972,416	21,603	5.44

Musa scaffold after GapCloser (step6)

Gaps within the 1,532 scaffolds were filled with the GapCloser program using the 330 b pair-end illumina sequencing libraries (50x) generated to correct the first version of the banana *Musa acuminata* reference genome. Of the total of 27,691 gap regions, 9,838 were closed.

Efficiency of the assembly process

The final assembly (Table 2) was composed of 1,532 scaffolds and showed a cumulated size of 450.7 Mb corresponding to 86% of the estimated size of the DH-Pahang genome. Ninety percent of the assembly was in 267 scaffolds and the N50 was 3.0 Mb. Gaps in scaffolds represent only 45.2 Mb (10.0% of the assembly).

Our semi-automatic pipeline applied on *Musa* scaffold correction allowed significant improvement of the assembly. The first step, by adding HCPSL data to the scaffolding process, decreased scaffold number from 7,513 to 2,267 and raised the N50 from 1.3 Mb to 1.5 Mb. Interestingly, the core of our user dependent pipeline (steps 2 to 4) decreased again the scaffold number of 30% (2,267 to 1,572) and had a much more critical effect on the N50 value which raises from 1.5 Mb to 2.9 Mb (Table 2 and supplemental Figure S6).

In comparison to the initial assembly of (D'Hont et al., 2012), the cumulated size of the new assembly is reduced by 21.5 Mb. This reduction is mainly due to the insertion of small scaffolds into previous gaps of larger scaffolds. The total size of the assembly, lower than expected, can be explained at least for a part by difficulties in assembling correctly the repeated fraction of the genome regions (45S and 5S ribosomal DNA, transposons, retro-transposons and tandem repeats). These repeat-rich sequences are often collapsed into single regions, resulting in a reduced size for the total assembly (Hahn et al., 2014). For example, 10.6 Mb rDNA have been found in the unassembled reads of DH-Pahang (D'Hont et al., 2012). In addition, 12 mitochondrial scaffolds were identified using BLAST (blastn, e-value 1e-100) of mitochondrial coding sequences of *Phoenix dactylifera* (NC_016740) (Fang et al., 2012) against assembled scaffolds. These mitochondrial scaffolds, presenting a cumulated size of 7.2 Mb, were also removed from the final nuclear assembly.

Musa pseudo-molecule improvement (step 7)

Genotyping data were used to group, order and orient scaffolds relative to each other into 11 pseudo-molecules. Of the 23,430 selected markers, 23,249 could be grouped using JoinMap and 21,851 mapped to a unique position on the 387 scaffold final version. Among them, 248 markers were discarded since they create local discrepancies in scaffolds clearly associated

Table 4 : Statistics on *Musa acuminata* pseudomolecule assembly beween first version and version 2

	Version1						Version2					
	scaffold cumulated size	Nb	scaffold N50	Nb	N in scaffolds	%	scaffold cumulated size	Nb	scaffold N50	Nb	N in scaffolds	%
chr01	27,571,529	22	2,245,470	4	3,459,727	12.5	29,067,552	30	1,394,891	2	2,151,480	7.4
chr02	22,052,597	22	1,755,924	3	2,961,122	13.4	29,509,134	27	2,676,329	3	3,555,070	12.0
chr03	30,468,307	22	3,785,391	3	3,981,002	13.1	35,017,413	31	9,733,574	2	2,329,119	6.7
chr04	30,050,316	13	8,856,836	2	3,343,441	11.1	37,104,143	17	7,838,899	3	2,076,824	5.6
chr05	29,375,369	21	2,773,165	4	3,488,635	11.9	41,848,132	52	2,239,696	5	3,976,084	9.5
chr06	34,896,279	30	7,330,853	2	4,472,335	12.8	37,589,864	36	9,841,105	2	2,328,163	6.2
chr07	28,615,304	22	5,244,634	3	4,262,894	14.9	35,025,021	31	6,378,715	3	4,518,654	12.9
chr08	35,437,139	27	2,556,008	3	5,002,970	14.1	44,883,571	57	9,906,416	2	3,821,170	8.5
chr09	34,145,263	37	1,544,587	6	5,397,793	15.8	41,302,925	39	2,119,922	3	3,398,494	8.2
chr10	33,662,572	33	1,266,487	5	5,753,963	17.1	37,671,811	31	1,798,308	3	3,318,350	8.8
chr11	25,512,624	15	7,530,813	2	2,838,651	11.1	27,952,850	16	7,787,879	2	1,979,175	7.1
mitochondrion	-	-	-	-	-	-	7,218,240	12	616,199	4	37,503	0.5
chrUn_random	140,423,018	7,249	183,001	168	36,766,009	26.2	46,507,017	1,153	137,262	86	11,685,573	25.1

with a linkage group by a majority of markers. Markers located on small scaffolds that were not split at step 1 and for which no linkage group majority could be found were also discarded and these scaffolds were not grouped in the pseudo-molecules. The remaining 21,603 markers were used to order and orient 376 scaffolds into the 11 pseudo-molecules (Figure 4). This step was performed with an average of 5.44 markers per 100 kb (Table 3) using methods and tools described in *Step7*.

Finally, a total of 397 Mb have been anchored, representing 89.5 % of the nuclear assembly (*versus* 70 % in version 1) and including all scaffolds larger than 1 Mb. Each pseudo-molecule comprised from 16 to 57 scaffolds and N50 in pseudo-molecules varied between 1.4 Mb to 9.9 Mb. The mean gap proportion varied from 5.6% to 12.9% in pseudo-molecules and was of 25.1% in unanchored scaffolds (Table 4).

In comparison to the first pseudo-molecule assembly version, we corrected the position of relatively few scaffolds from one pseudo-molecule to another (Figure 5, Supplemental Figure S7). One major change concerned three scaffolds that were previously attributed to chromosome 1 and that the new scaffolding and new genetic marker data now link together in chromosome 4. These regions of chromosomes 1 and 4 displayed massive segregation distortion that created pseudo-linkage between markers from these two regions and hindered the first version of the anchoring (D'Hont et al., 2012).

Most of the scaffolds that were newly integrated to the pseudo-molecules or were relocated on the same pseudo-molecule belonged to peri-centromeric regions. Due to the relatively low number of genetic markers (652 compared to 21,603) used to perform the anchoring in the first assembly, many small scaffolds had no marker and could not be anchored or had only one marker and could not be orientated. However, even with the new mapping data, some scaffolds in these peri-centromeric regions remain with an approximate order. Indeed, due to a high proportion of repeated sequences, these regions are assembled into small scaffolds and the density of genetic markers that mapped on these regions is low (Figure 4). In addition, the recombination rate is low (or even suppressed) in these peri-centromeric regions (Figure 4) (Chen et al., 2002; Gill et al., 1996; Hall et al., 2003; Wu et al., 2003). This low recombination rate in addition to genotyping errors can lead to marker miss-ordering and ultimately to scaffold miss-ordering. The method proposed in this paper is designed to take into account these genotyping errors using groups of markers already ordered into blocks corresponding to scaffolds, but for scaffolds with very few genetic markers and/or low recombination the ordering remains approximate.

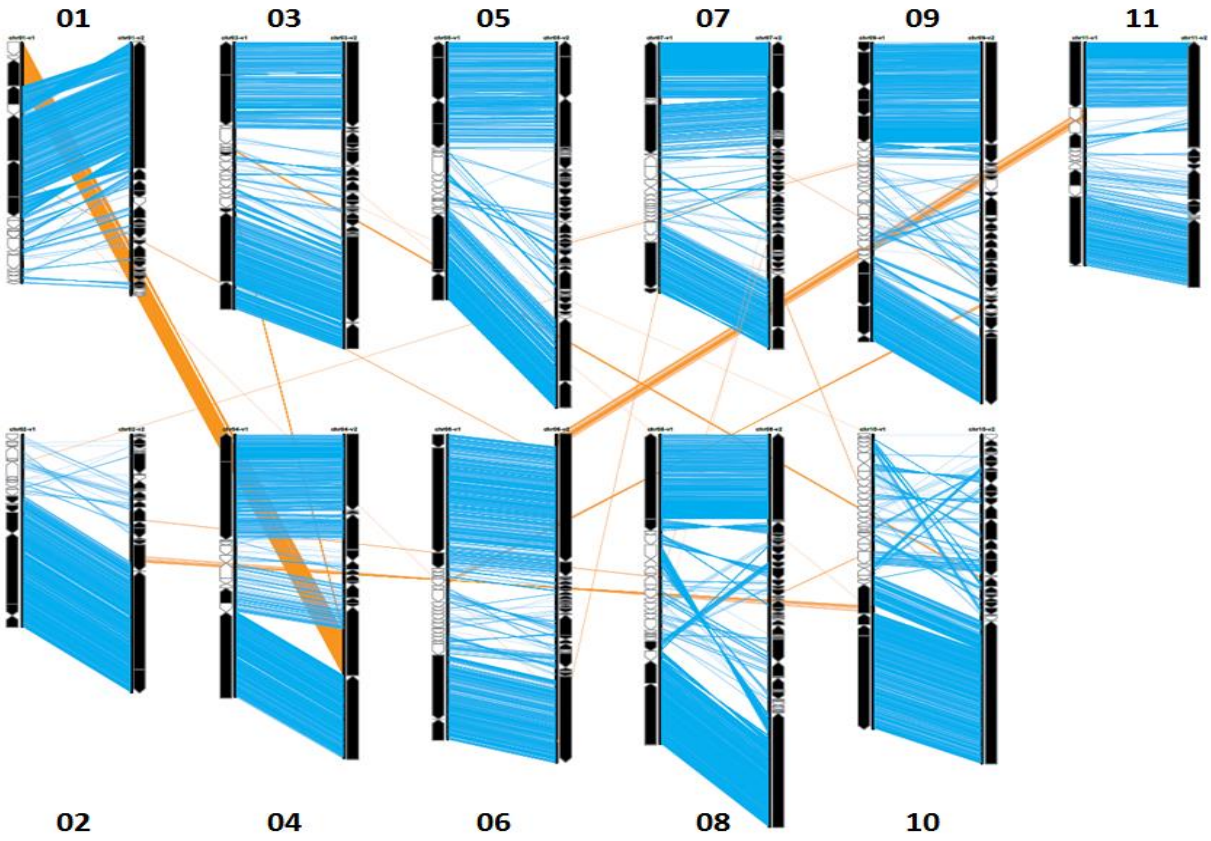


Figure 5: Comparison of scaffold anchoring between first version and new version of *Musa acuminata* pseudo-molecules. For each chromosomes, the new pseudo-molecule assembly (right) is compared to the pseudo-molecule assembly of D'Hont et al (2012) (left). Colored links joined identical markers mapped along the two assemblies. Blue: markers located on same pseudo-molecules between the two assemblies. Orange: markers located on different pseudo-molecules between the two assemblies. Boxes symbolize scaffolds and their orientation along the pseudo-molecules.

Table 5: Statistics on annotation transfert between the first release of the assembly and the new release.

	First release (D'hont et al 2012)			New release		
identifier	Size (bp)	Nb	%	Size (bp)	Nb	%
chr01	27,573,629	2,832	7.76	29,070,452	2,427	6.71
chr02	22,054,697	2,328	6.38	29,511,734	2,566	7.10
chr03	30,470,407	3,244	8.89	35,020,413	3,443	9.52
chr04	30,051,516	3,366	9.23	37,105,743	4,123	11.40
chr05	29,377,369	2,953	8.10	41,853,232	3,264	9.03
chr06	34,899,179	3,694	10.13	37,593,364	4,004	11.08
chr07	28,617,404	2,765	7.58	35,028,021	2,909	8.05
chr08	35,439,739	3,440	9.43	44,889,171	3,622	10.02
chr09	34,148,863	3,116	8.54	41,306,725	3,317	9.18
chr10	33,665,772	3,142	8.61	37,674,811	3,228	8.93
chr11	25,514,024	2,673	7.33	27,954,350	2,611	7.22
chrUn_random	141,147,818	2,926	8.02	46,622,217	542	1.50
Mitochondrial	N/A	N/A	N/A	7,218,240	96	0.27
Total	472,960,417	36,479	100	450,848,473	36,152	100

Annotation transfer

M. acuminata transcripts (D'Hont et al., 2012) were aligned to the genome using Exonerate v. 2.2 with the cdna2genome model and a maximum allowed intron size of 30 kb. Of the 36,479 predicted genes as of 19 September 2014, 36,056 (98.8%) genes were transferred from the first assembly version to the new assembly version (Table 5). Of the total number of transferred genes, 542 (1.5%) were located in unanchored scaffolds compared to 2,926 genes (8%) in the first version. Moreover, 96 genes were transferred onto the mitochondrial scaffolds.

Conclusion

In this paper, we describe a method and new tools to improve draft genome assembly with Next Generation Sequencing data. This method can intervene at different steps of the assembly process. The method comprised existing and newly developed bioinformatic tools that are arranged in a pipeline. Inside the pipeline, steps are independent and adapted to available datasets. They were implemented to the well-known easy to use GALAXY system and can be freely accessed through our bioinformatic platform Southgreen.

The first assembly of the banana genome (D'Hont et al., 2012) was improved by adding HCPSL, GBS data and optical mapping data to the original dataset. HCPSL data decreased the number of scaffolds contained in the assembly, by simply running again the scaffolding process but N50 was not affected at this step. HCPSL and/or GBS data from a mapping population were combined with *Scaffremodeler* bioinformatic tools i) to detect miss-assembled regions and identify scaffold fusions and junctions that automated scaffolding programs may have missed, ii) to propose figures of paired-read spanning candidate zones for expert manual validation, iii) to correct automatically validated zones and generate figures for final validation. This step allowed to reduce again the scaffold number and has a critical effect on N50 of the assembly which raised from 1.5 Mb to 3Mb. Optical mapping data validated junctions of scaffolds and slightly decreased the number of scaffolds to the final number. A final step of closing residual gaps was performed with standard NGS paired data.

In addition, we provide *Scaffhunter* tools to orient and order scaffolds and produce pseudo-molecules based on GBS genotyping data. These tools are based on a novel approach that exploits physical position of markers onto scaffolds. This approach was successfully adapted to the banana genome sequence. *Scaffhunter* tools were capable of partially resolving the problem of genotyping errors contained in the dataset and the high level of segregation distortion.

The *Musa* nuclear genome final assembly is composed of 1,520 scaffolds with a N50 of 3 Mb, 10% of unknown positions and 89.5% of sequence anchored of 11 pseudo-molecules. The improvements we made on the banana reference genome sequence will have great impact on future analysis based on this now improved high-quality reference sequence.

Availability and requirements

Operating system: All major Linux platforms

Programming language: python

Requirement: python2.6, perl, blastall, bwa, bowtie, bowtie2, java, picard-tools, samtools, circos5.1

Warning some versions of bwa (bwa-0.7.9a) may not be compatible with samtools.

Datasets (contigs, scaffold assembly and Pseudo-molecules) will be available through the banana genome hub (http://banana-genome.cirad.fr/) (Droc et al., 2013)

5 kb library will be deposited in NCBI Sequence Read Archive

Acknowledgments

The authors thanks the DArT company and CRP-RTB for funding DArTSeq data. We also thank the SouthGreen Bioinformatics Platform – UMR AGAP - CIRAD (http://southgreen.cirad.fr) for providing us with computational resources. We thank Daniele Roques from CRB and Christophe Jenny for providing 'Pahang' leaves and the 'Pahang' segregating population from the CIRAD research station in Guadeloupe, French West Indies.

Bibliography

Alkan, C., Sajjadian, S., and Eichler, E.E. (2011). Limitations of next-generation genome sequence assembly. Nat. Methods 8, 61–65.

Boetzer, M., and Pirovano, W. (2012). Toward almost closed genomes with GapFiller. Genome Biol. 13, R56.

Boetzer, M., and Pirovano, W. (2014). SSPACE-LongRead: scaffolding bacterial draft genomes using long read sequence information. BMC Bioinformatics 15, 211.

Boetzer, M., Henkel, C.V., Jansen, H.J., Butler, D., and Pirovano, W. (2011). Scaffolding pre-assembled contigs using SSPACE. Bioinformatics 27, 578–579.

Bolger, M.E., Weisshaar, B., Scholz, U., Stein, N., Usadel, B., and Mayer, K.F. (2014). Plant genome sequencing — applications for crop improvement. Curr. Opin. Biotechnol. 26, 31–37.

Chain, P.S.G., Grafham, D.V., Fulton, R.S., FitzGerald, M.G., Hostetler, J., Muzny, D., Ali, J., Birren, B., Bruce, D.C., Buhay, C., et al. (2009). Genome Project Standards in a New Era of Sequencing. Science 326, 236–237.

Chen, M., Presting, G., Barbazuk, W.B., Goicoechea, J.L., Blackmon, B., Fang, G., Kim, H., Frisch, D., Yu, Y., Sun, S., et al. (2002). An Integrated Physical and Genetic Map of the Rice Genome. Plant Cell Online 14, 537–545.

Cruz, V.M. (2013). Molecular Genetic Characterization of Lesquerella New Industrial Crop Using DArTseq Markers. In Plant and Animal Genome XXI Conference, (Plant and Animal Genome).

Dayarian, A., Michael, T., and Sengupta, A. (2010). SOPRA: Scaffolding algorithm for paired reads via statistical optimization. BMC Bioinformatics 11, 345.

D'Hont, A., Denoeud, F., Aury, J.-M., Baurens, F.-C., Carreel, F., Garsmeur, O., Noel, B., Bocs, S., Droc, G., Rouard, M., et al. (2012). The banana (Musa acuminata) genome and the evolution of monocotyledonous plants. Nature 488, 213–217.

Dong, Y., Xie, M., Jiang, Y., Xiao, N., Du, X., Zhang, W., Tosser-Klopp, G., Wang, J., Yang, S., Liang, J., et al. (2013). Sequencing and automated whole-genome optical mapping of the genome of a domestic goat (*Capra hircus*). Nat. Biotechnol. 31, 135–141.

Donmez, N., and Brudno, M. (2013). SCARPA: scaffolding reads with practical algorithms. Bioinformatics 29, 428–434.

Droc, G., Larivière, D., Guignon, V., Yahiaoui, N., This, D., Garsmeur, O., Dereeper, A., Hamelin, C., Argout, X., Dufayard, J.-F., et al. (2013). The Banana Genome Hub. Database 2013.

Fang, Y., Wu, H., Zhang, T., Yang, M., Yin, Y., Pan, L., Yu, X., Zhang, X., Hu, S., Al-Mssallem, I.S., et al. (2012). A Complete Sequence and Transcriptomic Analyses of Date Palm (*Phoenix dactylifera* L.) Mitochondrial Genome. PLoS ONE 7, e37164.

Feuillet, C., Leach, J.E., Rogers, J., Schnable, P.S., and Eversole, K. (2011). Crop genome sequencing: lessons and rationales. Trends Plant Sci. 16, 77–88.

Gao, S., Sung, W.-K., and Nagarajan, N. (2011). Opera: reconstructing optimal genomic scaffolds with high-throughput paired-end sequences. J. Comput. Biol. 18, 1681–1691.

Gill, K.S., Gill, B.S., Endo, T.R., and Taylor, T. (1996). Identification and high-density mapping of gene-rich regions in chromosome group 1 of wheat. Genetics 144, 1883–1891.

Gritsenko, A.A., Nijkamp, J.F., Reinders, M.J.T., and Ridder, D. de (2012). GRASS: a generic algorithm for scaffolding next-generation sequencing assemblies. Bioinformatics 28, 1429–1437.

Hahn, M.W., Zhang, S.V., and Moyle, L.C. (2014). Sequencing, Assembling, and Correcting Draft Genomes Using Recombinant Populations. G3 GenesGenomesGenetics.

Hall, S.E., Kettler, G., and Preuss, D. (2003). Centromere Satellites From Arabidopsis Populations: Maintenance of Conserved and Variable Domains. Genome Res. 13, 195–205.

Kejnovsky, E., Hawkins, J., and Feschotte, C. (2012). Plant Transposable Elements: Biology and Evolution. In Plant Genome Diversity Volume 1, J.F. Wendel, J. Greilhuber, J. Dolezel, and I.J. Leitch, eds. (Springer Vienna), pp. 17–34.

Krzywinski, M., Schein, J., Birol, İ., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J., and Marra, M.A. (2009). Circos: An information aesthetic for comparative genomics. Genome Res. 19, 1639–1645.

Levy-Sakin, M., and Ebenstein, Y. (2013). Beyond sequencing: optical mapping of DNA in the age of nanotechnology and nanoscopy. Curr. Opin. Biotechnol. 24, 690–698.

Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows–Wheeler transform. Bioinformatics 26, 589–595.

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., He, G., Chen, Y., Pan, Q., Liu, Y., et al. (2012). SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. GigaScience 1, 18.

Mardis, E.R. (2011). A decade's perspective on DNA sequencing technology. Nature 470, 198–203.

Mascher, M., and Stein, N. (2014). Genetic anchoring of whole-genome shotgun assemblies. Front. Genet. 5.

Mascher, M., Muehlbauer, G.J., Rokhsar, D.S., Chapman, J., Schmutz, J., Barry, K., Muñoz-Amatriaín, M., Close, T.J., Wise, R.P., Schulman, A.H., et al. (2013). Anchoring and ordering NGS contig assemblies by population sequencing (POPSEQ). Plant J. 76, 718–727.

Michael, T.P., and Jackson, S. (2013). The First 50 Plant Genomes. Plant Genome 6, -.

Neely, R.K., Deen, J., and Hofkens, J. (2011). Optical mapping of DNA: Single-molecule-based methods for mapping genomes. Biopolymers 95, 298–311.

Van Ooijen, J.W. (2011). Multipoint maximum likelihood mapping in a full-sib family of an outbreeding species. Genet. Res. 93, 343–349.

Pop, M., Kosack, D.S., and Salzberg, S.L. (2004). Hierarchical Scaffolding With Bambus. Genome Res. 14, 149–159.

Salmela, L., Mäkinen, V., Välimäki, N., Ylinen, J., and Ukkonen, E. (2011). Fast scaffolding with small independent mixed integer programs. Bioinformatics 27, 3259–3265.

Schatz, M., Witkowski, J., and McCombie, W.R. (2012). Current challenges in de novo plant genome sequencing and assembly. Genome Biol. 13, 243.

Slater, G., and Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics 6, 31.

Spindel, J., Wright, M., Chen, C., Cobb, J., Gage, J., Harrington, S., Lorieux, M., Ahmadi, N., and McCouch, S. (2013). Bridging the genotyping gap: using genotyping by sequencing

(GBS) to add high-density SNP markers and new value to traditional bi-parental mapping and breeding populations. Theor. Appl. Genet. 1–18.

Swain, M.T., Tsai, I.J., Assefa, S.A., Newbold, C., Berriman, M., and Otto, T.D. (2012). A post-assembly genome-improvement toolkit (PAGIT) to obtain annotated genomes from contigs. Nat. Protoc. 7, 1260–1284.

The International Wheat Genome Sequencing Consortium (IWGSC) (2014). A chromosome-based draft sequence of the hexaploid bread wheat (Triticum aestivum) genome. Science 345.

Vanneste, K., Maere, S., and Van de Peer, Y. (2014). Tangled up in two: a burst of genome duplications at the end of the Cretaceous and the consequences for plant evolution. Philos. Trans. R. Soc. B Biol. Sci. 369.

Williams, L.J.S., Tabbaa, D.G., Li, N., Berlin, A.M., Shea, T.P., MacCallum, I., Lawrence, M.S., Drier, Y., Getz, G., Young, S.K., et al. (2012). Paired-end sequencing of Fosmid libraries by Illumina. Genome Res. 22, 2241–2249.

Wu, J., Mizuno, H., Hayashi-Tsugane, M., Ito, Y., Chiden, Y., Fujisawa, M., Katagiri, S., Saji, S., Yoshiki, S., Karasawa, W., et al. (2003). Physical maps and recombination frequency of six rice chromosomes. Plant J. 36, 720–730.

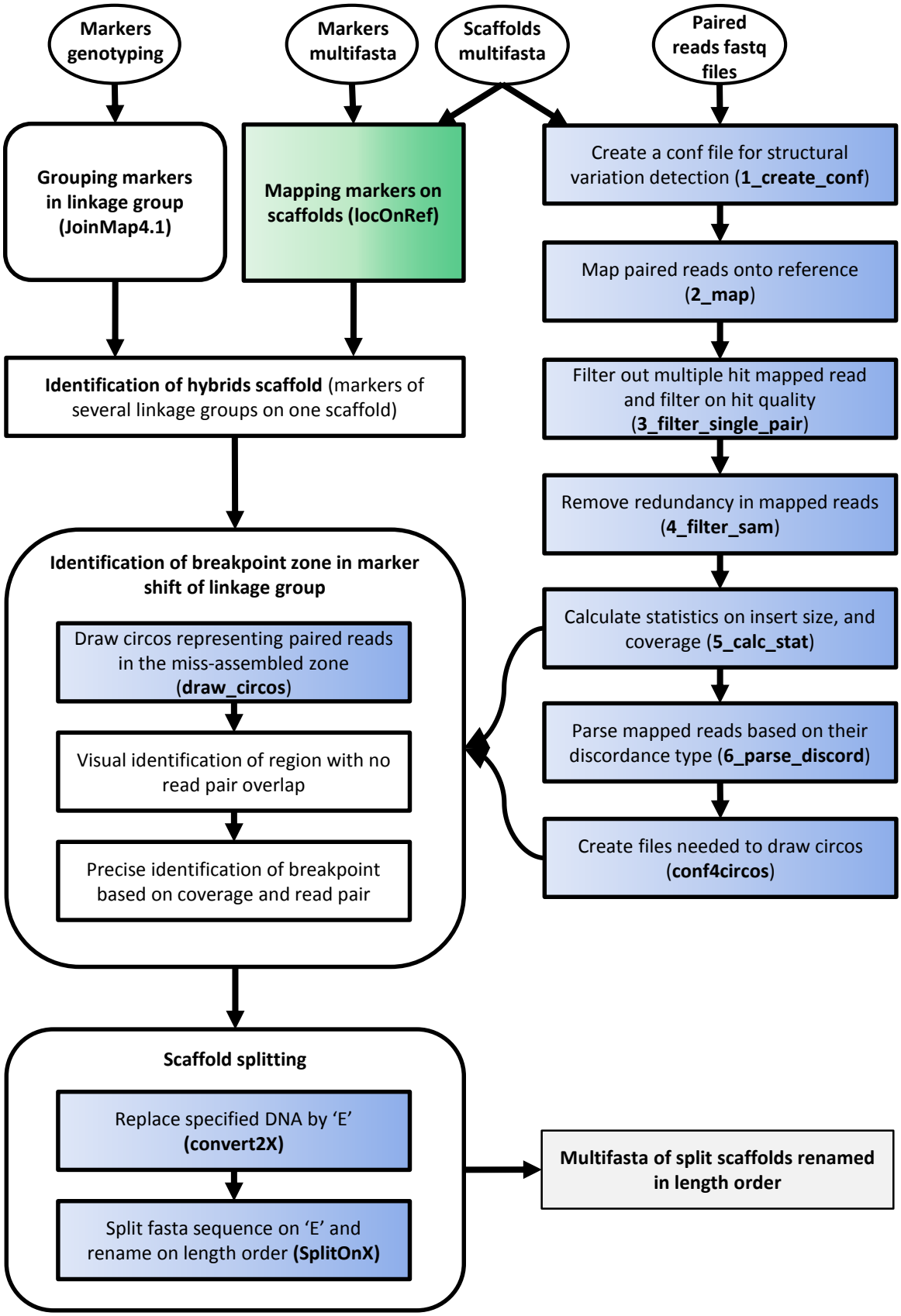


Figure S1: Overview of step2 that aims at identifying scaffold/contig miss-assemblies. Programs developed under Scaffhunter and Scaffremodler tools are in green and blue rectangles respectively. Grey rectangle presents final output. Program names are under brackets. For details, refer to Material and Methods section.

Supplementary Figures

Figure S1: Overview of step2 that aims at identifying scaffold/contig miss-assemblies.

Figure S2: Overview of step3 that aims at identifying and performing scaffold fusions and junctions.

Figure S3: Overview of step4 that aims at re-estimating gaps in scaffolds.

Figure S4: Overview of step7 that aims at ordering scaffolds into pseudo-molecules.

Figure S5: Example of scaffold fusion verification.

Figure S6: Evolution of scaffold number (red) and scaffold N50 (blue) in the assembly of the *Musa acuminata* reference sequence.

Figure S7: Dot plot of comparison of scaffold anchoring between first version (abscissa) and new version (ordinate).

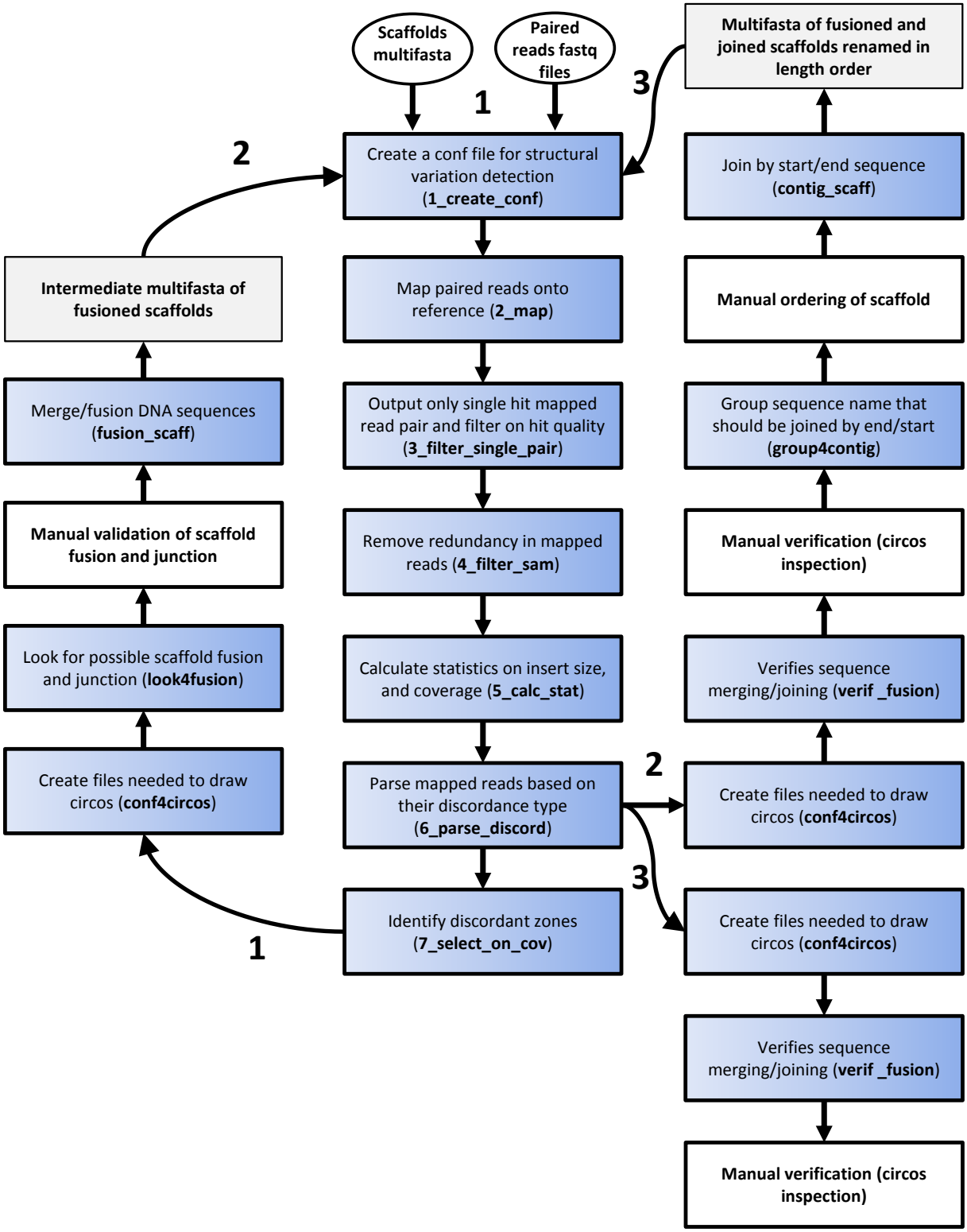


Figure S2: Overview of step3 that aims at identifying and performing scaffold fusions and junctions. Programs developed under Scaffremodler tools are in blue rectangles. Grey rectangles present intermediate and final outputs. Program names are under brackets. For details, refer to Material and Methods section.

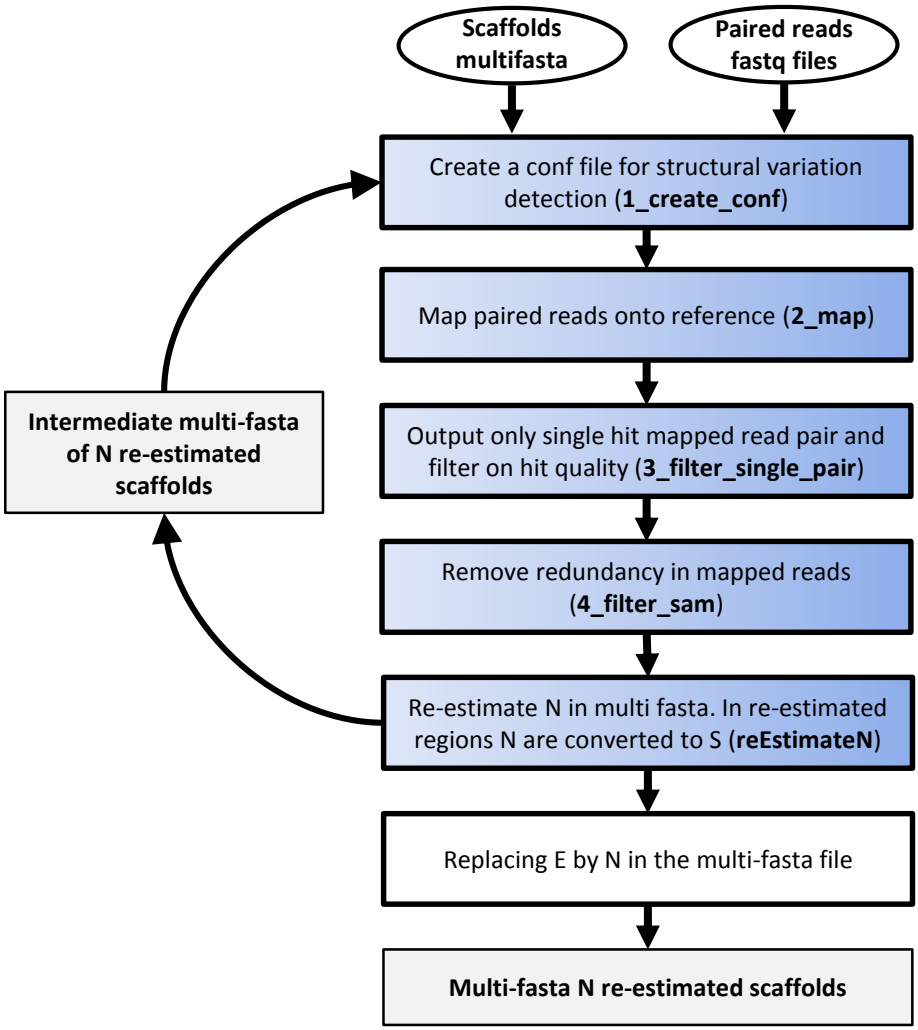


Figure S3: Overview of step4 that aims at re-estimating gaps in scaffolds. Programs developed under Scaffremodler tools are in blue rectangles. Grey rectangles present intermediate and final outputs. Program names are under brackets. For details, refer to Material and Methods section.

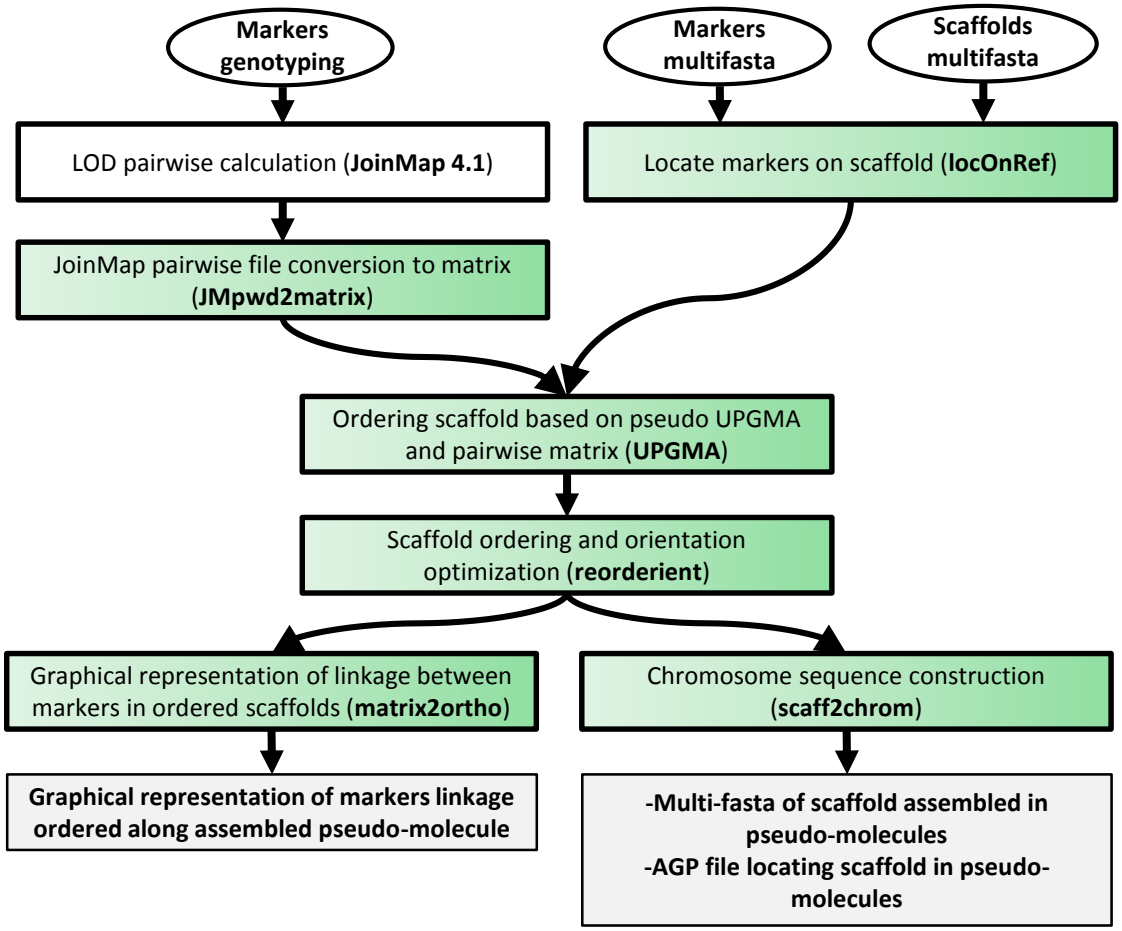


Figure S4: Overview of step7 that aims at ordering scaffolds into pseudo-molecules. Programs developed under Scaffhunter tools are in green rectangles. Final outputs are in grey rectangles. Program names are under brackets. For details, refer to Material and Methods section.

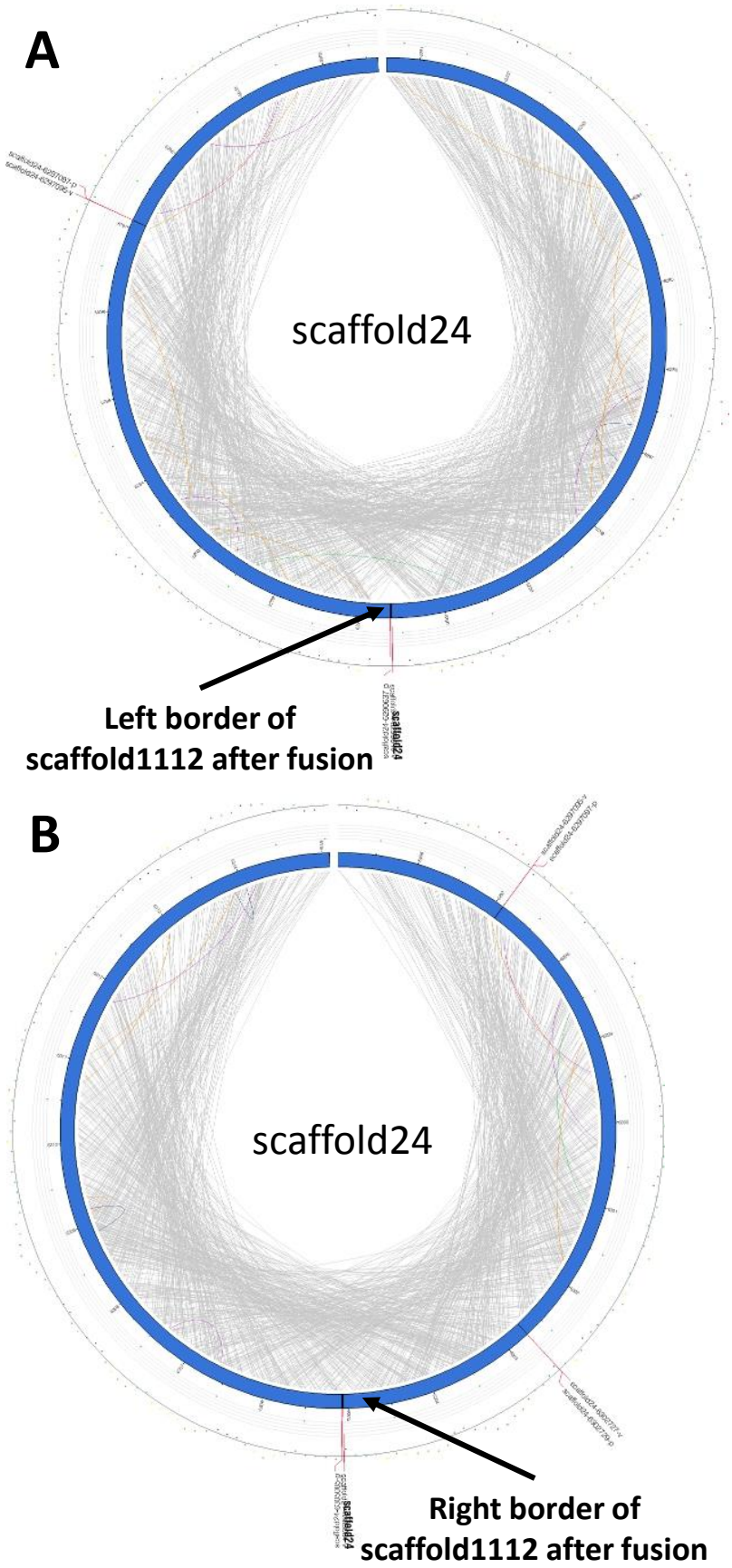


Figure S5: Example of scaffold fusion verification. Graphical representation of paired read mapping at the boundaries of the fusion of scaffold1112 into scaffold24. **(A)** Verification of the 5' fusion extremity. **(B)** Verification of the 3' fusion extremity. This representation is drawn using Scaffremodler tools. In the inner circle, read pair links are color coded according to their orientation and insert size: concordant pairs (correct orientation and insert size) are drawn in grey, discordant pair due to insert size are drawn in orange and red respectively for smaller and greater insert size. Pair showing the reverse-reverse orientation are drawn in purple, forward-forward orientation in green and pair having the complete reverse orientation relative to the expected (reverse-forward or forward-reverse, depending of paired library construction) are drawn in blue. The second circle represents the scaffolds and black regions locate gaps in the scaffold. The following circle is a scatter plot presenting the proportion of discordant reads on window size of 1 kb. The last circle is a scatter plot of read coverage on window size of 100 bases. Text locates start-1 (v) and end+1 (p) of gap regions in scaffold. Scale is expressed in 1 kb unit. The absence of discordant reads overlapping both 5' and 3' fusion zones validate the fusion performed. The presence of well orientated read pairs (grey link) overlapping scaffold fusion regions (left and right borders) confirm scaffold improvement performed.

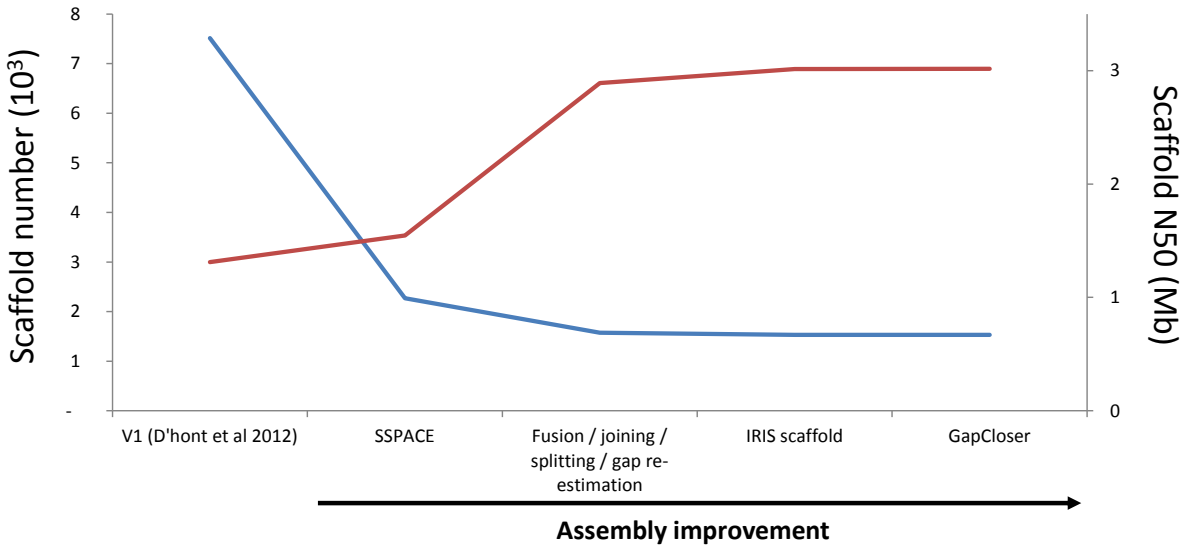


Figure S6: Evolution of scaffold number (red) and scaffold N50 (blue) in the assembly of the *Musa acuminata* reference sequence. The evolution of scaffold number and scaffold N50 during assembly improvement process is reported and compared to the first version of the assembly.

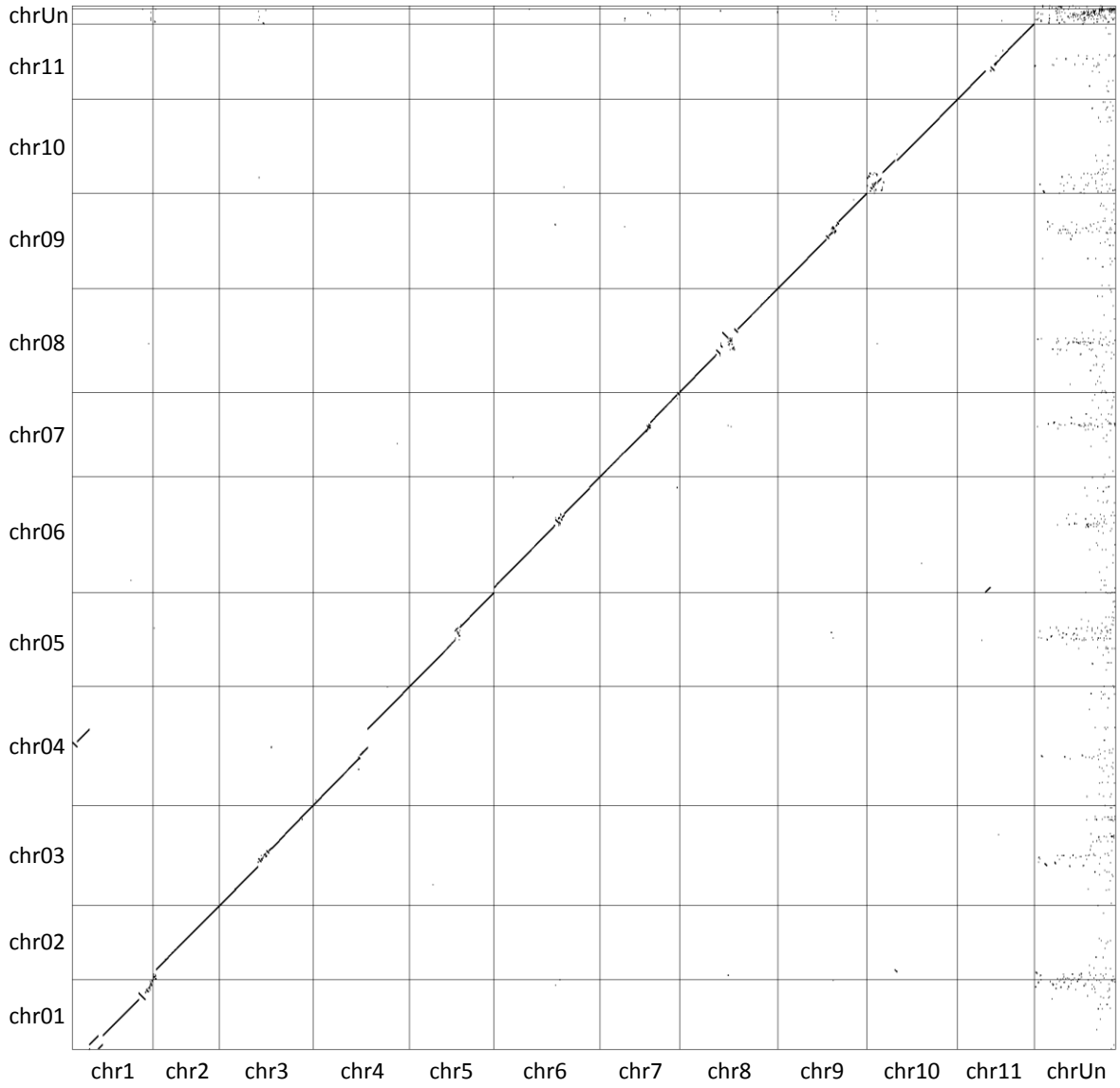


Figure S7: Dot plot of comparison of scaffold anchoring between first version (abscissa) and new version (ordinate). The dot-plot is made comparing gene location in first assembly version and second assembly version.

I.2 Assemblage du génome chloroplastique du bananier

Le chloroplaste est un organite spécifique aux plantes et aux algues. Il joue un rôle dans la synthèse des sucres, le stockage des réserves, la synthèse d'acides aminés, de lipides, de vitamines et de pigments. La taille de son génome varie entre 120 et 160 kb et présente en général une structure quadripartite comprenant une 'Large Single Copy region' (LSC) de 80-90 kb et une 'Small Single Copy region' (SSC) de 16-27 kb séparées par une région inversée répétée (IR) de 20-28 kb présente en deux copies (revue dans Chumley et al., 2006). Le génome chloroplastique présente une structure et un contenu en gènes bien conservés chez les angiospermes (Palmer, 1991; Raubeson and Jansen, 2005). Il code en général 4 ARNs ribosomaux, 30 ARNs messagers et 80 gènes codant des protéines. Cependant, le séquençage de nombreux génomes chloroplastiques a révélé la présence de réarrangements structuraux (Chumley et al., 2006; Martin et al., 2014; Palmer and Thompson, 1982), de variations de la taille des IR (Guisinger et al., 2010) ou encore de pertes de gènes chez certaines lignées (Martin et al., 2014; Millen et al., 2001).

Le génome chloroplastique du bananier a été assemblé à partir des séquences obtenues pour la production de la séquence du génome nucléaire du bananier. En effet, même si l'ADN utilisé pour réaliser le séquençage du génome nucléaire avait été enrichi en noyaux, il comprenait une certaine proportion d'ADN correspondant aux génomes chloroplastique et mitochondrial.

Les lectures (de type 454, http://www.454.com/) correspondant au génome chloroplastique ont été extraites du jeu de données par une recherche de similarité avec le génome chloroplastique d'une espèce proche, *Phoenix dactylifera* (Yang et al., 2010), en utilisant le programme BLAST (Altschul et al., 1990). Les lectures obtenues ont ensuite été assemblées en contigs. Ces contigs ont ensuite été allongés par une approche itérative de recherche de lectures, s'alignant aux extrémités des contigs dans le jeu de donnée initial suivi d'un ré-assemblage.

L'assemblage a montré que le génome chloroplastique du bananier est une molécule circulaire de 169 972 paires de bases présentant la structure quadripartite standard communément retrouvée chez les angiospermes. Dans cet article, nous avons confirmé la coexistence de deux structures chloroplastiques correspondant aux deux orientations possibles de la zone SSC par rapport à la zone LSC, en utilisant des séquences issues du re-séquençage d'extrémités de BACs. Une analyse de la structure de ce génome a révélé une extension de la zone inversée répétée par rapport à la SSC, se traduisant par une duplication complète de trois gènes et la

duplication partielle d'un quatrième gène. Le bananier est la seule monocotylédone où une telle expansion a été observée à ce jour.

Le séquençage du génome chloroplastique a également permis l'identification de répétitions de type microsatellites qui ont permis le développement de nouveaux marqueurs SSR. Ces marqueurs ont été testés pour leur potentiel de discrimination des sous-espèces de *Musa acuminata* et différentes accessions de *Musa.* Les résultats montrent que ces SSR sont plus discriminants que les marqueurs chloroplastiques disponibles jusque-là et donc qu'ils seront utiles pour mieux comprendre l'origine des bananiers cultivés.

Ce travail est présenté en détail dans l'article qui suit intitulé **The Complete Chloroplast Genome of Banana (Musa acuminata, Zingiberales)- Insight into Plastid Monocotyledon Evolution** publié dans la revue PLoS One.

Publication n°2

Title

The complete chloroplast genome of banana (*Musa acuminata*, Zingiberales): insight into plastid monocotyledons evolution

Guillaume Martin[1], Franc-Christophe Baurens[1], Céline Cardi[1], Jean-Marc Aury[2], Angélique D'Hont[1*]

1. CIRAD (Centre de coopération Internationale en Recherche Agronomique pour le Développement), UMR AGAP, F-34398 Montpellier, France

2. Genoscope, 2 rue Gaston Crémieux, BP5706, 91057 Evry, France.

* : Corresponding author

OPEN ACCESS Freely available online

PLOS ONE

The Complete Chloroplast Genome of Banana (*Musa acuminata*, Zingiberales): Insight into Plastid Monocotyledon Evolution

Guillaume Martin[1], Franc-Christophe Baurens[1], Céline Cardi[1], Jean-Marc Aury[2], Angélique D'Hont[1]*
1 CIRAD (Centre de coopération Internationale en Recherche Agronomique pour le Développement), UMR AGAP, Montpellier, France, 2 Genoscope, Evry, France

Abstract

Background:Banana (genus *Musa*) is a crop of major economic importance worldwide. It is a monocotyledonous member of the Zingiberales, a sister group of the widely studied Poales. Most cultivated bananas are natural *Musa* inter-(sub-)specific triploid hybrids. A *Musa acuminata* reference nuclear genome sequence was recently produced based on sequencing of genomic DNA enriched in nucleus.

Methodology/PrincipalFindings: The *Musa acuminata* chloroplast genome was assembled with chloroplast reads extracted from whole-genome-shotgun sequence data. The *Musa* chloroplast genome is a circular molecule of 169,972 bp with a quadripartite structure containing two single copy regions, a Large Single Copy region (LSC, 88,338 bp) and a Small Single Copy region (SSC, 10,768 bp) separated by Inverted Repeat regions (IRs, 35,433 bp). Two forms of the chloroplast genome relative to the orientation of SSC versus LSC were found. The *Musa* chloroplast genome shows an extreme IR expansion at the IR/SSC boundary relative to the most common structures found in angiosperms. This expansion consists of the integration of three additional complete genes (*rps15*, *ndhH* and *ycf1*) and part of the *ndhA* gene. No such expansion has been observed in monocots so far. Simple Sequence Repeats were identified in the *Musa* chloroplast genome and a new set of *Musa* chloroplastic markers was designed.

Conclusion:The complete sequence of *M. acuminata* ssp *malaccensis* chloroplast we reported here is the first one for the Zingiberales order. As such it provides new insight in the evolution of the chloroplast of monocotyledons. In particular, it reinforces that IR/SSC expansion has occurred independently several times within monocotyledons. The discovery of new polymorphic markers within *Musa* chloroplast opens new perspectives to better understand the origin of cultivated triploid bananas.

Citation: Martin G, Baurens F-C, Cardi C, Aury J-M, D'Hont A (2013) The Complete Chloroplast Genome of Banana (*Musa acuminata*, Zingiberales): Insight into Plastid Monocotyledon Evolution. PLoS ONE 8(6): e67350. doi:10.1371/journal.pone.0067350

Editor: James G. Umen, Donald Danforth Plant Science Center, United States of America

Received January 22, 2013; Accepted May 16, 2013; Published June 28, 2013

Funding: Centre de coopération Internationale en Recherche Agronomique pour le Développement (CIRAD), French National Research Agency (ANR) and Commissariat a` l'Energie Atomique (CEA). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: angelique.d'hont@cirad.fr

Introduction

Chloroplasts are the photosynthetic organelles that provide energy for plants and algae. They are also involved in major functions such as sugar synthesis, starch storage, the production of several amino acids, lipids, vitamins and pigments and also in key sulfur and nitrogen metabolic pathways. In angiosperms, chloroplastic (cp) genomes exist at least in part as a circular DNA molecule [1] ranging from 120 to 160 kb in length. Most cp genomes have a quadripartite organization comprising two copies of 20 to 28 kb Inverted Repeats (IRs) which separate the rest of the genome into a 80–90 kb Large Single Copy region (LSC) and a 16–27 kb Small Single Copy region (SSC) [2]. In angiosperms, the cp genome usually encodes 4 rRNAs, 30 tRNAs, and about 80 unique proteins. Earlier studies, using restriction site mapping, have demonstrated that gene content, gene order, and genome organization are largely conserved within land plants [3,4]. However, with the increasing number of whole cp genome available, many structural rearrangements, large IR expansion and gene loss have been reported [2,5,6]. These events can be used for the reconstruction of plant phylogeny [7]. Besides, the availability of whole chloroplast genomes or complete sets of cp genes have helped resolving relationships among major clades of angiosperms [8,9] with more accuracy than even well-chosen "Lucky Genes" [10]. Most of the reported complete monocotyledons chloroplast genomes are from the Poales group (so far 31 of the 46 complete chloroplast genomes deposited in Genbank). It is thus important to have more representatives of other clades to better understand the evolution of cp genome within monocots.

Bananas (genus *Musa*, family *Musa*ceae) are monocotyledons from the Zingiberales, a sister group of the Poales. Banana is of major economic importance in many tropical and subtropical countries where it is vital for food security and also a major source of incomes. Bananas are widely exported to industrialized countries where they represent the most popular fruit. A reference sequence of *Musa acuminata* nuclear genome has recently been published based on sequencing a DNA extract enriched in nucleus [11], yet providing additional sequence data to assemble a chloroplastic

genome. In banana, the peculiar paternal inheritance of the mitochondrial genome associated to the classic maternal inheritance of the chloroplast genome [12] make cytoplasmic markers potentially very useful for analyzing the origin of cultivars, most of which are spontaneous triploid inter-(sub)-specifc hybrids [13–15]. In previous studies based on RFLP [16] or PCR-RFLP [17] a total of nine different chloroplastic patterns have been identified among cultivated bananas and related wild species. However, most *M. acuminata* sub-species and cultivars had identical pattern restraining the identification of cultivars progenitors [18].

In this study, we report the assembly, annotation and structure analysis of the complete cp genome of banana. We compare its organization (gene content, IR expansion/contraction, structural rearrangement) with the complete genome of 34 monocots and 10 more basal angiosperms. We also provided new cp markers designed from Simple Sequence Repeats (SSR).

Materials and Methods

Sequence Data

A reference nuclear genome sequence of the doubled-haploid Pahang accession (DH-Pahang) was produced based on DNA extraction enriched in nuclear content. A total of 27,495,411 reads were generated using Roche/454 GSFLX pyrosequencing platfor*M.* An addition of 1,069,954 paired-Sanger 10 kb insert-size reads and 49,216 paired-Sanger BAC-ends sequenced on two BAC libraries generated with HindIII and BamHI restriction enzymes were produced [11].

The plastid reads were extracted from the total using blast similarity search against *Phoenix dactylifera* whole chloroplast genome (NC_013991). The 454 filtered reads were then assembled into sequence contigs using de novo assembly with Newbler. A total of six contigs were obtained. Using a python script, an iterative elongation for both ends of each contig using the total 454 reads was applied to ensure that contribution of *Musa* specific sequences was taken into account. The resulting four contigs (one contig for each region except two for the IRs) were then ordered based on P. *dactylifera* chloroplast structure. A mapping step using the pairedSanger 10 kb insert-size reads was then applied to confirm and correct contig junctions. A total of 1,800,008 GS FLX Titanium reads and 33,583 paired-Sanger reads were mapped to the assembled plastid genome representing 6.5% and 3.1% of the total 454 and Sanger reads for an average coverage of 5,341 X (sd =2,048), the large standard deviation mainly due to the doubling of coverage in the IRs. The minimum coverage was 619 X and the maximum coverage was reached in the IRs with a value of 9,500 X. The junction between the two contigs corresponding to the IRs was confirmed with the Sanger reads. The four junctions between the single-copy regions and IRs were confirmed by PCR.

LSC Orientation Relative to SSC

In order to verify the orientation of the SSC region relative to the LSC region, paired-Sanger BAC-end reads were mapped on the assembled *Musa* chloroplast genome using BLAST. Only pairs presenting more than 90% identity on more than 60% of their length were conserved. An additional filter was applied to conserve only pairs having a mate on the SSC region while the other was on the LSC. A total of 180 paired-Sanger BAC-ends were retained. Orientation visualization of the different paired BAC-ends reads was performed using CIRCOS [19] and was used to infer LSC orientation relative to SSC.

Genome Annotation

The genome was annotated by using DOGMA [20], followed with manual corrections for start codons. Intron positions were determined based on those of *P. dactylifera* [21] and *Elaeis guineensis* [22]. The transfer RNA genes were annotated using DOGMA and tRNAscan-SE (version1.23) [23]. Some intron-containing genes in which exons are too short to be detected were identified based on comparisons to corresponding exons in *P. dactylifera* and *E. guineensis*. The resulting annotated sequence has been deposited at the European Nucleotide Archive under accession number HF677508.

Codon Usage

Codon usage frequencies and the relative synonymous codon usage (RSCU) was calculated from coding sequences (CDS) of all different protein coding genes in the *M. acuminata* chloroplast genome using seqinr R-cran package [24].

Cp DNA Transfers to the Nucleus

Chloroplast DNA transfers to the nucleus were detected using Blast based approach. The assembled *M. acuminata* chloroplast genome, with one of its IR removed, was compared to the 11 chromosomes of *Musa* nuclear reference genome with high stringency blast parameter (e-value ,10^{25}, hit length .100 bp). A per base insertion value of each plastid base has been calculated as described in The Tomato Genome Consortium [25].

Phylogenetic Analysis

The phylogeny was performed using 79 plastid protein-coding genes derived from 48 plant species (Table S1) with complete chloroplast sequence, most belonging to monocotyledons. A codon based alignment was performed for each gene using homemade scripts that grouped together homologous genes and then converted them into proteins. An alignment was then applied to the protein sequence using MAFFT [26] and this protein alignment was then used to make the codon based aligment. Each aligned gene was then concatenated into a single matrix. Missing genes were replaced by Ns. A nucleotide matrix of 76,524 sites was then constituted. Evolutionary model choice was performed using jModelTest 2.0.2 software [27]. A maximum likelihood (ML) phylogenetic analysis was then performed using GTR+G+I model of sequence evolution using PhyML v3.0 [28]. Branch support was estimated based on aLRT statistics.

Musa Chloroplast Structure Comparison with others Whole cp Genomes

Gene positions of the different cp genomes were collected from the Genbank file and ordered based on their positions within the genome. Gene order and composition were then compared between the different species. Large events, e.g. gene loss, IR gene gain/loss, large structural rearrangement, relative to the basal angiosperm Amborella trichopoda [29] were recorded and used to infer scenarios in the different monocot lineages.

Short Tandem Repeats

Microsatellites (mono-, di-, tri-, tetra-, penta-, and hexanucleotide repeats) detection was performed using MISA [30] with minimum number of repeats of 10, 5, 4, 3, 3, 3 for 1, 2, 3, 4, 5, 6 unit size respectively (Table S2). Minisatellites (unit size $10) were detected manually using dot plot with Gepard software [31] with the *Musa* chloroplast sequence plotted against itself. Sequences with unit repeat equal to or higher than 10 bp repeated tandemly at least twice were conserved. The dot plot was

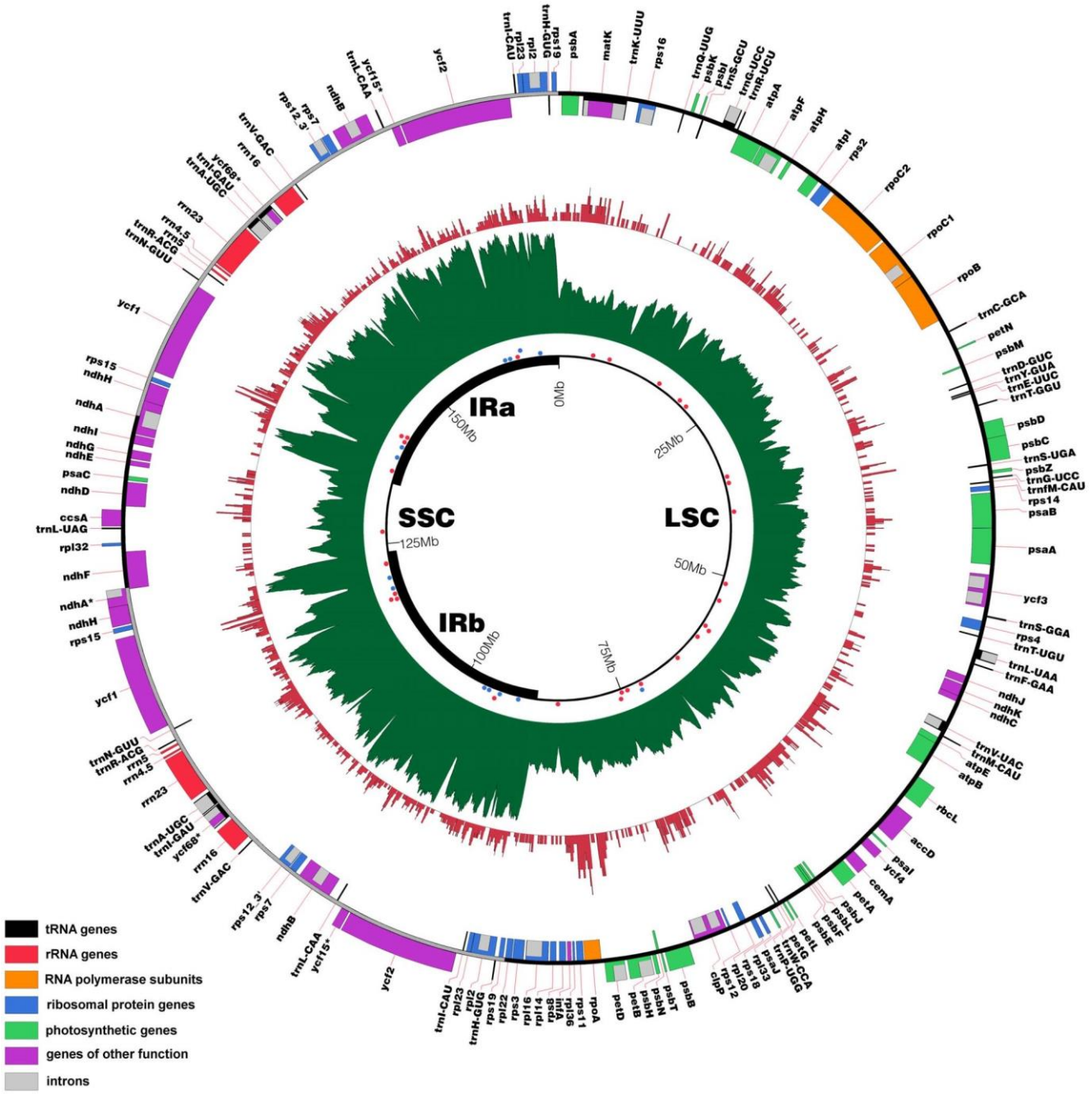


Figure 1. Circular *Musaacuminata* chloroplast map. Genes are represented with boxes inside or outside the circle to indicate clockwise or counterclockwise transcription direction respectively. The color of the gene boxes indicates the functional group to which the gene belongs. Read depth of the genome is represented in the inner green circle. The locations of the short tandem repeats, tested for their polymorphism, are represented with red and blue dots for microsatellites and minisatellites respectively. The per base insertion value in the nucleus is drawn in the red circle. The per base insertion value of the IR analyzed has been divided by two and applied to both IR. Pseudogenes are marked with asterisks. doi:10.1371/journal.pone.0067350.g001

A total of 32 SSR located all over the plastid genome (Figure 1) were tested for their polymorphism in a testing panel comprising *M.* boman, *M.* balbisiana and three *M. acuminata* spp., *M.*a. ssp banksii, *M.*a. ssp zebrina and *M.*a. ssp malaccensis (DH-Pahang) with the Applied BiosystemsH 35006L Genetic Analyzer. This set included 7 minisatellites and 26 microsatellites. The 12 most polymorphic markers were evaluated onto 5 additional cultivated accessions, including the triploid accession Cavendish Grande Naine, all belonging to the *Musa* chloroplastic group II [16].

Table *Musa* plastome

Plastome characteristics	
Size (bp)	169,972
LSC size in bp (%)	88,338 (52.0)
SSC size in bp (%)	10,768 (6.3)
IR length in bp	35,433
Size in bp (%) coding regions	100,277 (59.0)
Size in bp (%) of protein-encoding regions	88,336 (52.0)
Size in bp (%) of introns	19,312 (11.4)
Size in bp (%) of rRNA	9,056 (5.3)
Size in bp (%) of tRNA	2,885 (1.7)
Size in bp (%) of IGS	50,389 (29.6)
Number of different genes	113
Number of different protein-encoding genes	79
Number of different tRNA genes	30
Number of different rRNA genes	4
Number of different genes duplicated by IR	24
Number of different genes with introns	18
Overall % GC content	36.8
% GC content in protein-encoding regions	37.3
% GC content in introns	37.8
% GC content in IGS	31.6
% GC content in rRNA	55.2
% GC content in tRNA	53.1

doi:10.1371/journal.pone.0067350.t001

Results and Discussion

General Feature of *Musa acuminata* cp Genome

The *M. acuminata* chloroplast genome is a DNA molecule of 169,972 bp in length. Similar to most other angiosperms, the chloroplast genome of *M. acuminata* is circular with a quadripartite structure: a pair of Inverted Repeats (IRs) (35,433 bp) separated by the Single Copy region (SSC) (10,768 bp) and Large Single Copy region (LSC) (88,338 bp) (Figure 1). A total of 136 functional genes were predicted, including 113 distinct genes comprising 79 protein-coding genes, 30 transfer RNA (tRNA) genes and 4 ribosomal RNA (rRNA) genes (Table 1). All 4 rRNA genes, 8 tRNA and 10 protein-coding genes are repeated in the IR. Protein-coding genes, tRNA and rRNA represent respectively 52.0%, 1.7% and 5.3% of the plastid genome. Non-coding DNA, including intergenic spacers (IGSs) and introns represent 41.0% of the genome. Similar to other plastid genomes, the overall GC content of the *M. acuminata* plastid genome is 36.8%. This value is slightly higher in protein coding genes (37.3%) and introns (37.8%), slightly lower in IGS (31.6%) while tRNA and rRNA show higher GC value with 53.1% and 55.2% respectively.

A total of 23,199 codons represent the 79 different proteincoding genes of the *M. acuminata* chloroplast genome. Among these, 2,350 (10.6%) code for leucine and 269 (1.2%) for cysteine, which are the most frequent and the least frequent amino acids, respectively (Table 2). The 30 different tRNA found in the chloroplast genome correspond to 28 different codons, at least one for each amino acid. Only 7 of the 28 different anticodon tRNAs encoded in the *Musa* plastid genome correspond to the most common codon (where synonymous codons exist). The codon usage is biased towards a high representation of A and T at the third position, as observed in most land plant chloroplast genomes [33].

The *M. acuminata* chloroplast genome has 18 different introncontaining genes, six of which are tRNA. Most have a single intron except two genes, clpP and ycf3, which contain two introns. The gene rps12 is trans-spliced and has the 59exon in the LSC and two exons in the IR. The *ycf1*5 and ycf68 genes were found to have 5 and 7 internal stop codons respectively. This suggests that *ycf1*5 and ycf68 have become pseudogenes in *M. acuminata* chloroplast genome. These two pseudogenes were mentioned in very few chloroplast studies and thus were not used in our phylogenetic study. The incomplete duplication of the 59 end of *ndhA* at the IRa and SSC boundary resulted in two *ndhA* gene copies: a pseudogene at the boundary of IRb and SSC and a complete copy at the IRa and SSC boundary.

LSC Orientation Relative to SSC

Due to the inverted repeated regions it was not possible to conclude on the orientation of the SSC relative to the LSC using the 454 and Sanger reads 10 kb paired reads. BAC-end-sequences (BES) were used to orient the SSC relative to LSC. A total of 77 BES, 29 and 48 in the Forward/Reverse and Reverse/Forward orientation respectively, support the orientation presented in this paper (Figure 2A). Another set of 103 BES, 29 and 74 in the Forward/Forward and Reverse/Reverse support a SSC in the reverse complement order (Figure 2B). These results imply that the two forms co-exist in the *M. acuminata* chloroplast genome. This coexistence of two orientation-forms has previously been reported in Phaseolus vulgaris [34] and Zea mays [35] using RFLP analysis.

DNA Transfer to the Nucleus

A total of 563 hits (Table 3) of more than 100 bp were found on the eleven chromosomes of *M. acuminata* for a cumulative length of 134,491 bp of the *Musa* nuclear genome (0.41 %). This value is situated between those of *Arabidopsis thaliana* (0.17%) [36] and tomato (0.75%) [25]. A much higher value (1.85%) had been found in the rice genome [37]. Matsuo et al. [38], using the rice nuclear genome, reported that the plant nuclear genome is in equilibrium between integration and elimination of the chloroplast genome. The various proportions of inserted chloroplast genome observed in plant species reveal different levels of equilibriu*M.* These variations may result from distinct speed of cp DNA transfer flow to the nucleus or a distinct speed of elimination of the inserted cp DNA or a combination of these two processes.

Based on a per base insertion value calculated for each plastid base, we showed that the cp DNA inserted in the *Musa* nuclear DNA originate from every part of the chloroplast genome and covers 57.4% of the chloroplast (without IRa) (Figure 1). The highest per base insertion values appeared around the regions carrying the rpoA gene and to a further extent in a region containing the *ycf1* gene. In the tomato genome, the per base insertion value was also higher in two regions carrying ycf genes [25].

Unlike tomato and rice that countain numerous large insertions, only 6 hits of more than 1 kb but not exceeding 2 kb were found on the *M. acuminata* nuclear genome. Chloroplast insertions were found on all chromosomes (Table 3) with a relatively homogeneous distribution unlike the uneven distribution observed in rice [37]. Chromosome 2, with 2.43% of all chloroplast genome insertions, was the chromosome with the least plastid insertion while chromosome 6 was the one having the most abundant plastid insertion (11.08%). The cp DNA insertions into the nuclear genome of *Musa* were evenly distributed over the chromosomes with a reduced number of insertions in pericentromeric regions

Table 2. Codon usage and codon-anticodon recognition pattern of the *Musa acuminata* chloroplast genome.

Amino acid	Codon	Number	RSCU[a]	Frequency[b]	Amino acid	Codon	Number	RSCU[a]	Frequency[b]
F	TTT	839	1.28	64.19	A	GCA	353	1.13	28.35
F	TTC	468	0.72	35.81	A	GCG	122	0.39	9.80
L	TTA	750	1.91	31.91	Y	TAT	682	1.58	78.75
L	TTG	497	1.27	21.15	Y	TAC	184	0.42	21.25
L	CTT	465	1.19	19.79	H	CAT	432	1.55	77.56
L	CTC	158	0.40	6.72	H	CAC	125	0.45	22.44
L	CTA	324	0.83	13.79	Q	CAA	640	1.54	77.11
L	CTG	156	0.40	6.64	Q	CAG	190	0.46	22.89
I	ATT	980	1.47	48.98	N	AAT	850	1.55	77.27
I	ATC	384	0.58	19.19	N	AAC	250	0.45	22.73
I	ATA	637	0.96	31.83	K	AAA	907	1.51	75.52
M	ATG	546	1.00	100.00	K	AAG	294	0.49	24.48
V	GTT	468	1.42	35.51	D	GAT	763	1.61	80.32
V	GTC	168	0.51	12.75	D	GAC	187	0.39	19.68
V	GTA	505	1.53	38.32	E	GAA	976	1.50	74.90
V	GTG	177	0.54	13.43	E	GAG	327	0.50	25.10
S	TCT	512	1.72	28.72	C	TGT	200	1.49	74.35
S	TCC	290	0.98	16.26	C	TGC	69	0.51	25.65
S	TCA	342	1.15	19.18	W	TGG	394	1.00	100.00
S	TCG	161	0.54	9.03	R	CGT	326	1.39	23.24
S	AGT	390	1.31	21.87	R	CGC	76	0.33	5.42
S	AGC	88	0.30	4.94	R	CGA	312	1.33	22.24
P	CCT	374	1.57	39.33	R	CGG	105	0.45	7.48
P	CCC	196	0.82	20.61	R	AGA	441	1.89	31.43
P	CCA	280	1.18	29.44	R	AGG	143	0.61	10.19
P	CCG	101	0.42	10.62	G	GGT	539	1.38	34.46
T	ACT	454	1.54	38.41	G	GGC	154	0.39	9.85
T	ACC	235	0.80	19.88	G	GGA	643	1.64	41.11
T	ACA	366	1.24	30.96	G	GGG	228	0.58	14.58
T	ACG	127	0.43	10.74	*	TAA	41	1.56	51.90
A	GCT	575	1.85	46.18	*	TAG	20	0.76	25.32
A	GCC	195	0.63	15.66	*	TGA	18	0.68	22.78

[a]: relative synonymous codon usage. [b]: codon frequency relative to each amino acid.

Codons shown in bold complement the anticodons of the tRNAs encoded in the chloroplast genome. Frequencies shown in bold indicate the most common codon(where synonymous codons exist for that amino acid or termination). doi:10.1371/journal.pone.0067350.t002

(Figure S2). However, this may be due to lower assembly quality of this type of regions.

Phylogenetic Analysis

A ML phylogenetic analysis was conducted based on 79 protein coding gene from 48 plant taxa. The resulting topology is presented in Figure 3 and Figure S1. All except two nodes are well supported with aLRT statistics higher than 0.98. The first not well supported node with a value of 0.972 is in the Basal Angiosperms group and is positioning *Chloranthus spicatus* at a basal position to the group constituted of *Drimys granadensis*, *Piper cenocladum*, *Calycanthus floridus*, *Magnolia kwangsiensis* and *Liriodendron tulipifera*. The second ambiguous node, with an aLRT value of 0.396, is the relative position of the Bamboo species *Ferrocalamus rimosivaginus* and *Acidosasa purpurea* at the basal position of the others Arundinarieae included in the analysis. Speciation of Zingiberales, Arecales and Poales has long been difficult to resolve and conflicting results have been reported [9]. Our *M. acuminata* chloroplast data positions Zingiberales as sister to the Poales. The Arecales is positioned as sister group to the Poales and

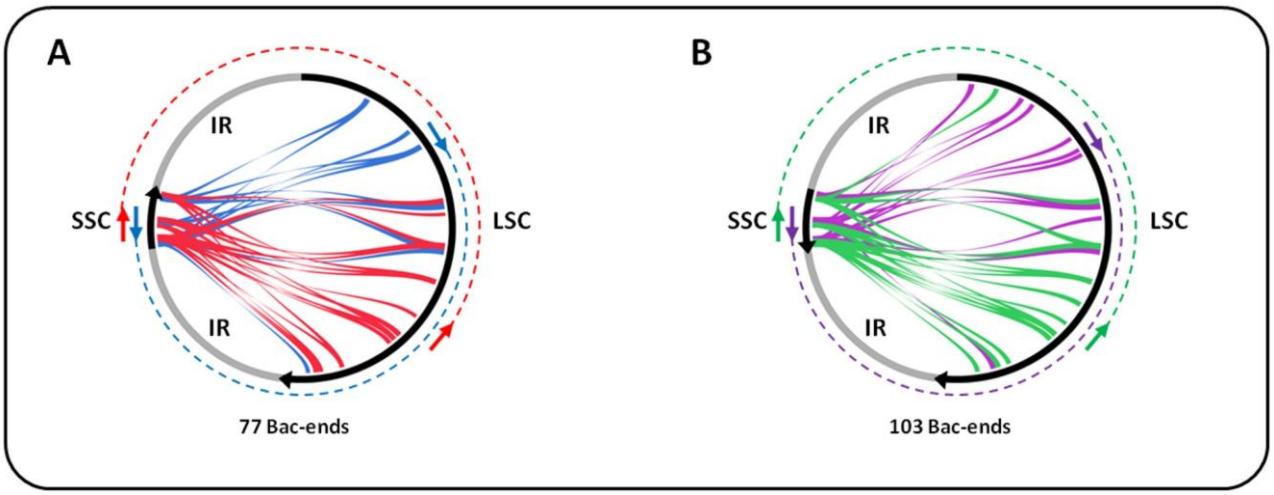


Figure 2. BAC-end-sequences (BESs) mapped on the LSC and SSC *Musa* chloroplast genome. A, BESs mapping with the Forward/Reverse (FR), and Reverse/Forward (RF) orientations respectively in blue and red, supporting the SSC orientation relative to the LSC as displayed in the assembled *Musa* chloroplast sequence. B, BESs mapping with the Forward/Forward (FF), and Reverse/Reverse (RR) orientations respectively in purple and green, supporting the presence of another form relative to the orientation of the SSC vs LSC in *M. acuminata*. doi:10.1371/journal.pone.0067350.g002

Zingiberales in agreement with previous study based on chloroplast genes [6,8,9]. However, these results differ from the phylogenetic trees obtained with 93 nuclear single genes, that regroup Zingiberales and Arecales in a sister group to the Poales [11]. Similar incongruence between analyses of single-copy nuclear genes and the chloroplast genes has been observed in the phylogenetic placement of the Malpighiales within the Rosids [39]. These incongruences may be caused by incomplete lineage sorting [40], long-branch attraction phenomenon [41,42] or chloroplast introgressions between *Musa*ceae and Poales ancestors (see [43] for example). Additional taxa sampling and coalescence-based analyses will be required to resolve this conflict.

Structural Comparison within Angiosperms

The *M. acuminata* chloroplast genome structure was compared to other angiosperms. Major chloroplast genome structural events (gene losses, IR expansion/contraction and structural rearrangements) and inferred scenarios impacting several Monocotyledons clades are reported on the phylogenetic analysis in Figure 3 (for details on each species and basal angiosperms see Figure S1).

Gene content. The *infA* gene has been lost through multiple independent events from at least 24 Angiosperm lineage chloroplast genomes [5]. It is present in the *Musa acuminata* chloroplast genome as well as other Monocotyledons studied so far to the exception of the Alismatale lineage (Figure 3 and [44]). The accD, *ycf1* and *ycf2* genes are annotated as functional genes in the *Musa* chloroplastic genome while they have been lost in Poaceae cp genomes [6]. In addition *clpP* and *rpoC1* introns, found in the *Musa* chloroplast genome, have been lost in Poaceae with the exception of the basal Poales Anomochloa [45]. The *AccD* and *ycf1* genes have also been lost in the Acoraceae and Orchidaceae. The *ndhB*, *ndhJ*, *ndhC*, *ndhK*, *ndhD*, *ndhF*, *ndhA*, *ndhH*, *ndhG* genes lost in all Orchidaceae genomes sequenced [46–48] are all annotated functional in the *Musa* chloroplastic genome as for the rps16 gene lost in Dioscorea elephantipes [49].

Structural rearrangements. Dot plot analysis showed that *Musa* chloroplastic genome organization is similar to those found within most angiosperms. The *M. acuminata* chloroplast genome does not present the major structural rearrangement of 30 kb found in Poaceae [6,50–53]. Relative to the *Musa* chloroplast genome, this rearrangement consisted in two inversions of 25 and 1 kb respectively and a translocation of about 5 kb all located in the same region.

IR expansion/contraction. The most derived chloroplast genome sequenced of Araceae, Bambusoideae, Poideae, Ehrhartoideae, Panicoideae and Anomochlooideae show events of IR/

Table 3. Chloroplast genome insertion into the nuclear genome in *Musa acuminata*.

Chromosomes	Nb_hits	Nb_bases	Proportion (%) of all cp insertions	Proportion (%) of the chromosome relative to the total nuclear DNA
chr01	57	10,468	5.49	8.32
chr02	24	4,637	2.43	6.65
chr03	44	9,629	5.05	9.19
chr04	63	17,752	9.31	9.06
chr05	39	6,993	3.67	8.86
chr06	83	21,129	11.08	10.53
chr07	42	10,838	5.68	8.63
chr08	44	10,482	5.50	10.69
chr09	48	11,125	5.83	10.30
chr10	60	13,659	7.16	10.16
chr11	59	17,779	9.32	7.70

doi:10.1371/journal.pone.0067350.t003

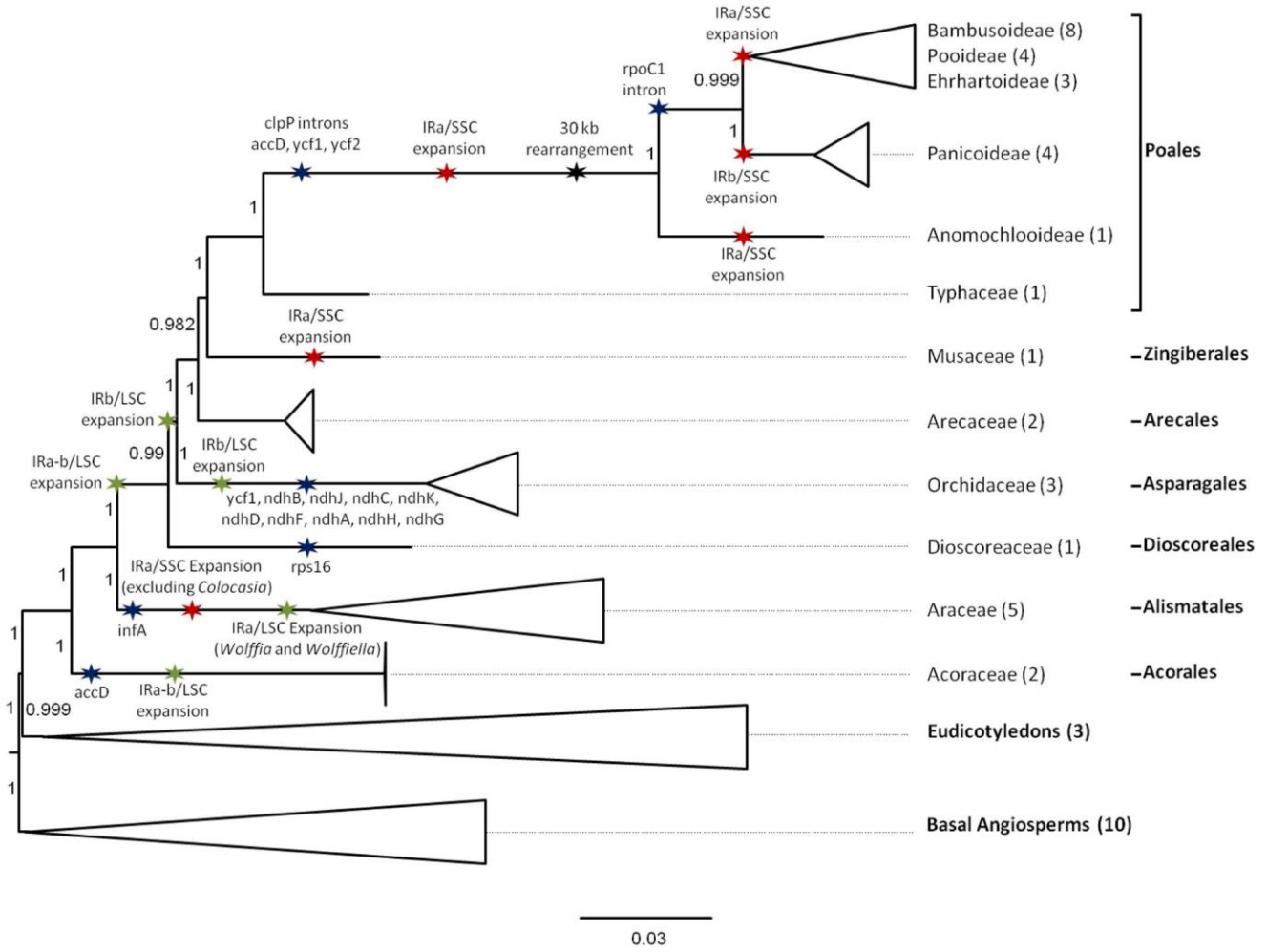


Figure 3. Condensed tree based on the maximum likelihood phylogenetic analysis constructed on 79 chloroplast protein coding genes of 10 basal angiosperms, 35 monocotyledons and 3 dicotyledons. The tree has a -lnL of 2527912.066159. Support values for ML are provided at the nodes. Gene losses in all members of the different clades are indicated with blue stars. Putative events of IR expansions/contractions in the monocots are indicated with red and green stars for IR/SSC and IR/LSC boundaries respectively. Major structural rearrangements are indicated with black stars. Numbers indicate aLRT branches support. doi:10.1371/journal.pone.0067350.g003

SSC expansion relative to Amborella trichopoda. IRs of the *Musa acuminata* chloroplast genome show an extreme extension that includes two additional genes (*rps15* and ndhH) plus the full sequence of *ycf1* and 1030 bp of the *ndhA* gene relative to the IR structure of Amborella trichopoda. The expansion is made at the IRa/ SSC junction and is the largest observed in monocots. In all other monocot groups where IR/SSC expansion is observed, except for the Panicoideae group, the expansion has occurred only at the IRa/SSC junction. The result of these expansions is the inclusion of the whole sequences of *ycf1* and *rps15* in the IRs and a part of the *ndhH* gene except for the Araceae group where *ndhH* is not always included. In the Panicoideae, the IRs contain *rsp15*, and a part of the *ndhF* gene suggesting that the IR/SSC extension has been made in two steps: first an IRa/SSC extension that has included *rps15* in the IR and a second step with an IRb/SSC extension including a part of the *ndhF* gene in the IR. These results suggest that an event of IRa/SSC extension has occurred prior to the divergence of Anomochlooideae, Bambusoideae, Poideae, Ehrhartoideae and Panicoideae including *rps15* gene in the IRs. After the divergence, Anomoclooideae group and Bambusoideae/ Poideae/Ehrhartoideae group have been subjected to independent additional IRa/SSC extension to include a part of *ndhH* in the IRs while Panicoideae have been subjected to IRb/SSC extension to include a part of the *ndhF* gene in the IRs. This scenario is similar to the one proposed by Guisinger et al. [6] but it adds the independents events of secondary IRa/SSC expansion in Anomochlooideae and the Bambusoideae/Pooideae/Ehrhatoideae group. This secondary IRa/SSC expansion provides further support for the sister relationship between Bambusoideae, Pooideae and Ehrhatoidea. The phylogenetic position of *M. acuminata* relative to Typhaceae at the basis of Poales and the chloroplastic structure of *Typha* showing no event of IR/SSC expansion suggest that *M. acuminata* has been subjected to an independent event of IRa/SSC expansion relative to Poales. Further investigation should be conducted to determine if this event is common to other *Musa*ceae and the Zingiberales. In the Araceae the most derived species show an IRa/SSC expansion while the basal species Colocasia esculenta and the sister group Acoraceae and the Dioscoreaceae, Orchidaceae and Arecaceae do not show IR/SSC expansion. This suggests that this event of IRa/ SSC expansion in the Araceae is

another independent event. To summarize, three major IRa/SSC expansions may have occurred

Table 4. Number of alleles detected within the Musaceae, Eumusa, *M. acuminata* ssp (M. a.) and within the chloroplastic group II samples.

Markers	Musaceae (10)	Eumusa (9)	M. a (8)	cp group II (6)
mMaClRcp01	4	4	4	3
mMaClRcp02	2	2	2	2
mMaClRcp19	4	3	2	1
mMaClRcp20	5	4	3	3
mMaClRcp25	4	4	4	4
mMaClRcp27	4	3	3	2
mMaClRcp29	4	3	2	1
mMaClRcp30	5	4	3	2
mMaClRcp31	3	3	2	1
mMaClRcp32	4	3	3	3
mMaClRcp33	5	5	5	4
mMaClRcp34	4	3	2	2
Average per marker	4.00	3.42	2.92	2.33

The number of accession tested for each group is in parenthesis.
doi:10.1371/journal.pone.0067350.t004

independently in monocotyledons, one in the Araceae, one in *Musa*ceae and one in the Poaceae. Three secondary independent events of IR/SSC expansion in the Poaceae have occurred, an IRa/SSC expansion in Anomochlooideae, an IRa/SSC expansion in Bambusoideae/Pooideae/Ehrhartoideae group and an IRb/ SSC expansion in Panicoideae.

All Monocots sequenced except the most basal Araceae show events of IR/LSC expansion relative to A. trichopoda (Figure 3). The Acoraceae and Dioscoreacea display the insertion of the trnH-GUG gene at the IRa/LSC boundary and a partial copy of the rps19 gene at the IRb/LSC boundary. The most derived plastid genomes sequenced of Araceae (Wolffia australiana and Wolffiella lingulata) only display a partial expansion of the IR including a partial copy of the rps19 gene at the IRb/LSC boundary. All sequenced plastid genomes belonging to the sister group of the Dioscoreacea (Poales, Zingiberales, Arecales and Asparagales) present the insertion of complete trnH-GUG and rps19 genes located in the LSC of Amborella at the IRa/LSC and IRb/LSC boundaries respectively. Asparagales show an additional IRb/LSC expansion as all their whole cp genome sequenced includes a partial copy of the rpl22 gene. The relative order of trnH-GUG and rps19 genes in IR suggests that in Acoraceae, Dioscoreaceae, Poales, Zingiberales, Arecales and Asparagales the expansion has been made in two steps as proposed by Mardanov et al. [44]: an IRa/LSC expansion leading to the inclusion of the trnH-GUG gene in the IR followed with an IRb/LSC expansion leading to the total or partial inclusion of the rps19 gene in the IR, depending of the group. The structure of the IR/LSC boundary observed in the different clades can be explained by three independent events of IRa-b/LSC expansion, one in Acoraceae, one in the most derived Araceae and one at the basis of the Dioscoreaceae, Poales, Zingiberales, Arecales and Asparagales group. A second round of IRb/LSC expansion has taken place in the last group excluding the Dioscoreaceae leading to the complete inclusion of the rps19 gene in the IR. A third round of expansion of IRb/LSC expansion has taken place in Asparagales leading to the partial inclusion of rpl22 gene in the IR as it has been proposed in Wang et al. [54].

Overview of the Short Tandem Repeats Landscape

Short tandem repeats (also named Simple sequence repeats (SSR)) can exhibit high variation within the same species and are thus considered valuable markers for population genetics [55,56] and phylogenetic analyses [57]. A total of 112 SSRs were detected in the *Musa* chloroplast genome. Among them, 54 are microsatellites (mono-, di-, tri-, tetra-, penta-, and hexanucleotide repeats) and 58 are minisatellites (unit size>10). Minisatellites detected have a unit repeat mean length of 20.8 bp with a minimum of 11 bp and a maximum of 43 bp. The most repeated minisatellite has 14 units of 30 bp repeated tandemly. Among the microsatellites, 39 are exclusively constituted of A/T nucleotides while only one microsatellite is exclusively constituted of C/G nucleotides. Fourteen microsatellites are a mixture of puric and pyrimidic bases.

Sixteen of the homopolymer loci contain multiple A or T nucleotides while only one contains multiple G or C nucleotides. This higher proportion of poly(A)/(T) relative to poly(G)/(C) has also been reported in Poceae [57] and more divergent species such as *Panax ginseng* and *Nicotiana tabacu*m [58], *Cucumis sativus* [59], *Magnolia kwangsiensis* [60], *Megaleranthis saniculifolia* [61] or *Sesamum indicum* [62]. However *P. ginseng* and *S. indicum* showed a slightly higher proportion of poly(G)/(C). In *Musa*, among the 10 dinucleotide repeat loci found, 8 are multiple AT or TA and 2 are multiple GA or AG. In Poaceae and *M. saniculifolia*, AT and TA repeats are the most common but others forms are found while only multiple AT or TA are reported in *S. indicuM.* In *Musa*, seven trinucleotide repeat loci, fourteen tetranucleotide, five pentanucleotide and one hexanucleotide are found. While in Poaceae tri-, tetra-, penta-, and hexanucleotide repeats are reported [57], no tetra-, penta- and hexanucleotide are reported in the eudicotyledon *S. indicum* and no hexanucleotide are reported in the eudicotyledon M. saniculifolia.

Musa Chloroplast PCR Markers

A total of 32 SSR (Table S3) were tested for their polymorphism within a sample of *Musa*. Seven markers appeared monomorphic and 25, 21 and 15 were polymorphic in *Musa*ceae, Eu*Musa* and within the *M. acuminata* sub-species, respectively. The 12 most polymorphic markers were further tested in a sample of six accessions belonging to the chloroplastic group II defined by Carreel et al. [16]. The number of haplotypes detected within our panel is presented in Table 4 for the 12 SSR markers. The average polymorphism level was 4.00, 3.42, 2.92 and 2.33 alleles per marker respectively in *Musa*ceae, Eu*Musa*, *M. acuminata* and within the chloroplastic group II. Among these 12 markers, 9 revealed polymorphism within the chloroplastic group II and showed from 2 to 4 alleles. This new set of chloroplastic PCR markers represents a new, fast and efficient tool for studying the diversity of bananas and the origin of cultivars. Most cultivated bananas are triploids derived from spontaneous hybridization between *M. acuminata* sub-species and a few other *Musa* species but their exact origin is still not completely understood [15,63]. Their high level of sterility complicates their use in breeding programs. In this context the identification of their fertile progenitors would be very useful for breeders.

Conclusion

We assembled, annotated and analyzed the complete chloroplast sequence of banana (*Musa acuminata* ssp malaccensis). This first Zingiberale chloroplast (cp) genome was compared to other available monocotyledon cp genomes, providing new insight in their evolution. IR/SSC expansion is particularly pronounced in

banana and has occurred independently several times within monocotyledons. The availability of new chloroplast markers within *Musa* opens new perspective to refine the phylogeny of *Musa* and the origin of cultivated triploid bananas.

Supporting Information

Figure S1 Maximum likelihood phylogenetic analysis based on 79 chloroplast protein coding genes of 45 basal angiosperms and monocotyledons and 3 Dicotyledons. The tree has a -lnL of 2527912.066159. Support values for ML are provided at the nodes. Gene losses in chloroplast genomes are indicated with red triangles. Green and red stars represent partial or total IR gain of genes belonging respectively to LSC or SSC relative to A. trichopoda structure. Green and red minus signs represent loss of one of the two partial or complete gene copies belonging to IR respectively to become member of LSC or SSC relative to A. trichopoda structure. (PDF)

Figure S2 Localization of cp DNA inserted in the nuclear genome of *M. acuminata*. (PDF)

Table S1Chloroplast genomes compared with the *M.acuminata* chloroplast. (PDF)

Table S2 Distribution of simple sequence repeats (SSRs) loci in the *M. acuminata* chloroplast genome.
(PDF)

Table S3 Makers, associated primer, and expected length tested for the polymorphism analysis. (PDF)

Acknowledgments

We thank the SouthGreen Bioinformatics Platform – UMR AGAP -
CIRAD (http://southgreen.cirad.fr) for providing us with computational resources. We thank Dr Jim Leebens-Mack for critical reading of the manuscript.

Author Contributions

Conceived and designed the experiments: GM FCB AD. Performed the experiments: GM FCB CC JMA. Analyzed the data: GM FCB. Wrote the paper: GM FCB AD.

References

1. Bendich AJ (2004) Circular Chloroplast Chromosomes: The Grand Illusion. ThePlant Cell Online 16: 1661–1666. doi:10.1105/tpc.160771.
2. Chumley TW, Palmer JD, Mower JP, Fourcade HM, Calie PJ, et al. (2006) TheComplete Chloroplast Genome Sequence of Pelargonium 6 hortorum: Organization and Evolution of the Largest and Most Highly Rearranged Chloroplast Genome of Land Plants. Molecular Biology and Evolution 23: 2175–2190. doi:10.1093/molbev/msl089.
3. Palmer JD (1991) Plastid chromosomes: structure and evolution. In: Bogorad L,Vasil I, editors. Cell Culture and Somatic Cell Genetics of Plants. San Diego: Academic Press. 5–53.
4. Raubeson LA, Jansen RK (2005) Chloroplast genomes of plants. In: Henry RJ,editor. Plant diversity and evolution: genotypic and phenotypic variation in higher plants. Cambridge: CAB International. 45–68.
5. Millen RS, Olmstead RG, Adams KL, Palmer JD, Lao NT, et al. (2001) ManyParallel Losses of infA from Chloroplast DNA during Angiosperm Evolution with Multiple Independent Transfers to the Nucleus. The Plant Cell Online 13: 645–658. doi:10.1105/tpc.13.3.645.
6. Guisinger M, Chumley T, Kuehl J, Boore J, Jansen R (2010) Implications of thePlastid Genome Sequence of Typha (Typhaceae, Poales) for Understanding Genome Evolution in Poaceae. J Mol Evol 70: 149–166. doi:10.1007/s00239009-9317-3.
7. Downie SR, Palmer JD (1992) Use of chloroplast DNA rearrangements inreconstructing plant phylogeny. In: Soltis PS, Soltis DE, Doyle JJ, editors. Molecular systematics of plants. New York: Chapman and Hall. 14–35.
8. Jansen RK, Cai Z, Raubeson LA, Daniell H, dePamphilis CW, et al. (2007) Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. Proceedings of the National Academy of Sciences 104: 19369–19374. doi:10.1073/ pnas.0709121104.
9. Givnish TJ, Ames M, McNeal JR, McKain MR, Steele PR, et al. (2010) Assembling the Tree of the Monocotyledons: Plastome Sequence Phylogeny and Evolution of Poales1. Annals of the Missouri Botanical Garden 97: 584–616. doi:10.3417/2010023.
10. Logacheva MD, Penin AA, Samigullin TH, Vallejo-Roman CM, Antonov AS (2007) Phylogeny of flowering plants by the chloroplast genome sequences: in search of a ‘‘lucky gene.’’ Biochemistry Moscow 72: 1324–1330. doi:10.1134/ S0006297907120061.
11. D'Hont A, Denoeud F, Aury J-M, Baurens F-C, Carreel F, et al. (2012) The banana (*Musa acuminata*) genome and the evolution of monocotyledonous plants. Nature 488: 213–217. doi:10.1038/nature11241.
12. Fauré S, Noyer J-L, Carreel F, Horry J-P, Bakry F, et al. (1994) Maternal inheritance of chloroplast genome and paternal inheritance of mitochondrial genome in bananas (*Musa acuminata*). Curr Genet 25: 265–269. doi:10.1007/ BF00357172.
13. Simmonds NW (1962) The evolution of the bananas. London: Longmans.
14. De Langhe E, Hrˇibova´ E, Carpentier S, Dolezˇel J, Swennen R (2010) Did backcrossing contribute to the origin of hybrid edible bananas? Annals of Botany 106: 849–857. doi:10.1093/aob/mcq187.
15. Perrier X, De Langhe E, Donohue M, Lentfer C, Vrydaghs L, et al. (2011) Multidisciplinary perspectives on banana (*Musa* spp.) domestication. Proceedings of the National Academy of Sciences 108: 11311–11318. doi:10.1073/ pnas.1102001108.
16. Carreel F, de Leon DG, Lagoda P, Lanaud C, Jenny C, et al. (2002) Ascertaining maternal and paternal lineage within *Musa* by chloroplast and mitochondrial DNA RFLP analyses. Genome 45: 679–692.
17. Boonruangrod R, Desai D, Fluch S, Berenyi M, Burg K (2008) Identification of cytoplasmic ancestor gene-pools of *Musa acuminata* Colla and *Musa* balbisiana Colla and their hybrids by chloroplast and mitochondrial haplotyping. Theor Appl Genet 118: 43–55. doi:10.1007/s00122-008-0875-3.
18. Lescot T (2011) The genetic diversity of the banana in figures. FruiTrop 189: 58–62.
19. Krzywinski M, Schein J, Birol I˙, Connors J, Gascoyne R, et al. (2009) Circos: An information aesthetic for comparative genomics. Genome Research 19: 1639–1645. doi:10.1101/gr.092759.109.
20. Wyman SK, Jansen RK, Boore JL (2004) Automatic annotation of organellar genomes with DOGMA.

Bioinformatics 20: 3252–3255. doi:10.1093/bioinformatics/bth352.
21. Yang M, Zhang X, Liu G, Yin Y, Chen K, et al. (2010) The Complete Chloroplast Genome Sequence of Date Palm (Phoenix dactylifera L.). PLoS ONE 5: e12762. doi:10.1371/journal.pone.0012762.
22. Uthaipaisanwong P, Chanprasert J, Shearman JR, Sangsrakru D, Yoocha T, et al. (2012) Characterization of the chloroplast genome sequence of oil palm (Elaeis guineensis Jacq.). Gene 500: 172–180. doi:10.1016/j.gene.2012.03.061.
23. Lowe TM, Eddy SR (1997) tRNAscan-SE: A Program for Improved Detection of Transfer RNA Genes in Genomic Sequence. Nucleic Acids Research 25: 0955–0964. doi:10.1093/nar/25.5.0955.
24. Charif D, Lobry J (2007) SeqinR 1.0–2: A Contributed Package to the R Project for Statistical Computing Devoted to Biological Sequences Retrieval and Analysis. In: Bastolla U, Porto M, Roman HE, Vendruscolo M, editors. Structural Approaches to Sequence Evolution. Biological and Medical Physics, Biomedical Engineering. Springer Berlin Heidelberg. 207–232. Available: http://dx.doi.org/10.1007/978-3-540-35306-5_10.
25. The Tomato Genome Consortium (2012) The tomato genome sequence provides insights into fleshy fruit evolution. Nature 485: 635–641. doi:10.1038/ nature11119.
26. Katoh K, Misawa K, Kuma K, Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transfor*M.* Nucleic Acids Research 30: 3059–3066. doi:10.1093/nar/gkf436.
27. Posada D (2008) jModelTest: Phylogenetic Model Averaging. Molecular Biology and Evolution 25: 1253–1256. doi:10.1093/molbev/msn083.
28. Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, et al. (2010) New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. Systematic Biology 59: 307–321. doi:10.1093/sysbio/syq010.
29. Goremykin VV, Hirsch-Ernst KI, Wo¨lfl S, Hellwig FH (2003) Analysis of the Amborella trichopoda Chloroplast Genome Sequence Suggests That Amborella Is Not a Basal Angiosper*M.* Molecular Biology and Evolution 20: 1499–1505. doi:10.1093/molbev/msg159.
30. Thiel T, Michalek W, Varshney R, Graner A (2003) Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (Hordeum vulgare L.). Theor Appl Genet 106: 411–422. doi:10.1007/s00122002-1031-0.
31. Krumsiek J, Arnold R, Rattei T (2007) Gepard: a rapid and sensitive tool for creating dotplots on genome scale. Bioinformatics 23: 1026–1028. doi:10.1093/ bioinformatics/btm039.
32. Rozen S, Skaletsky H (2000) Primer3 on the WWW for general users and for biologist programmers. In: Krawetz S, Misener S, editors. Bioinformatics Methods and Protocols in the series Methods in Molecular Biology. Totowa: Humana Press. 365–386.
33. Clegg MT, Gaut BS, Learn GH, Morton BR (1994) Rates and patterns of chloroplast DNA evolution. Proceedings of the National Academy of Sciences 91: 6795–6801.
34. Palmer JD (1983) Chloroplast DNA exists in two orientations. Nature 301: 92–
93. doi:10.1038/301092a0.
35. Oldenburg DJ, Bendich AJ (2004) Most Chloroplast DNA of Maize Seedlings in Linear Molecules with Defined Ends and Branched Forms. Journal of Molecular Biology 335: 953–970. doi:10.1016/j.jmb.2003.11.020.
36. Shahmuradov I, Akbarova Y, Solovyev V, Aliyev J (2003) Abundance of plastid DNA insertions in nuclear genomes of rice and Arabidopsis. Plant Mol Biol 52: 923–934. doi:10.1023/A:1025472709537.
37. Cullis CA, Vorster BJ, Van Der Vyver C, Kunert KJ (2009) Transfer of genetic material between the chloroplast and nucleus: how is it related to stress in plants? Annals of Botany 103: 625–633. doi:10.1093/aob/mcn173.
38. Matsuo M, Ito Y, Yamauchi R, Obokata J (2005) The Rice Nuclear Genome Continuously Integrates, Shuffles, and Eliminates the Chloroplast Genome to Cause Chloroplast–Nuclear DNA Flux. The Plant Cell Online 17: 665–675. doi:10.1105/tpc.104.027706.
39. Shulaev V, Sargent DJ, Crowhurst RN, Mockler TC, Folkerts O, et al. (2011) The genome of woodland strawberry (Fragaria vesca). Nat Genet 43: 109–116. doi:10.1038/ng.740.
40. Degnan JH, Rosenberg NA (2009) Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends in Ecology & Evolution 24: 332–340. doi:10.1016/j.tree.2009.01.009.
41. Hendy MD, Penny D (1989) A Framework for the Quantitative Study of Evolutionary Trees. Systematic Biology 38: 297–309. doi:10.2307/2992396.
42. Bergsten J (2005) A review of long-branch attraction. Cladistics 21: 163–193. doi:10.1111/j.1096-0031.2005.00059.x.
43. Renoult J, Kjellberg F, Grout C, Santoni S, Khadari B (2009) Cyto-nuclear discordance in the phylogeny of Ficus section Galoglychia and host shifts in plant-pollinator associations. BMC Evolutionary Biology 9: 248.
44. Mardanov A, Ravin N, Kuznetsov B, Samigullin T, Antonov A, et al. (2008) Complete Sequence of the Duckweed (Lemna minor) Chloroplast Genome: Structural Organization and Phylogenetic Relationships to Other Angiosperms. J Mol Evol 66: 555–564. doi:10.1007/s00239-008-9091-7.
45. Morris LM, Duvall MR (2010) The chloroplast genome of Anomochloa marantoidea (Anomochlooideae; Poaceae) comprises a mixture of grass-like and unique features. American Journal of Botany 97: 620–627. doi:10.3732/ ajb.0900226.
46. Chang C-C, Lin H-C, Lin I-P, Chow T-Y, Chen H-H, et al. (2006) The Chloroplast Genome of Phalaenopsis aphrodite (Orchidaceae): Comparative Analysis of Evolutionary Rate with that of Grasses and Its Phylogenetic Implications. Molecular Biology and Evolution 23: 279–291. doi:10.1093/ molbev/msj029.
47. Wu F-H, Chan M-T, Liao D-C, Hsu C-T, Lee Y-W, et al. (2010) Complete chloroplast genome of Oncidium Gower Ramsey and evaluation of molecular markers for identification and breeding in Oncidiinae. BMC Plant Biol 10: 1–12. doi:10.1186/1471-2229-10-68.
48. Jheng C-F, Chen T-C, Lin J-Y, Chen T-C, Wu W-L, et al. (2012) The comparative chloroplast genomic analysis of photosynthetic orchids and developing DNA markers to distinguish Phalaenopsis orchids. Plant Science 190: 62–73. doi:10.1016/j.plantsci.2012.04.001.
49. Hansen DR, Dastidar SG, Cai Z, Penaflor C, Kuehl JV, et al. (2007) Phylogenetic and evolutionary implications of complete chloroplast genome sequences of four early-diverging angiosperms: Buxus (Buxaceae), Chloranthus (Chloranthaceae), Dioscorea (Dioscoreaceae), and Illicium (Schisandraceae). Molecular Phylogenetics and Evolution 45: 547–563. doi:10.1016/j.ympev.2007.06.004.
50. Doyle JJ, Davis JI, Soreng RJ, Garvin D, Anderson MJ (1992) Chloroplast DNA inversions and the origin of the grass family (Poaceae). Proceedings of the

National Academy of Sciences 89: 7722–7726. doi:10.1073/pnas.89.16.7722.

51. Hiratsuka J, Shimada H, Whittier R, Ishibashi T, Sakamoto M, et al. (1989) The complete sequence of the rice (Oryza sativa) chloroplast genome: Intermolecular recombination between distinct tRNA genes accounts for a major plastid DNA inversion during the evolution of the cereals. Molec Gen Genet 217: 185–194. doi:10.1007/BF02464880.
52. Howe C, Barker R, Bowman C, Dyer T (1988) Common features of three inversions in wheat chloroplast DNA. Curr Genet 13: 343–349. doi:10.1007/ BF00424430.
53. Katayama H, Ogihara Y (1993) Structural alterations of the chloroplast genome found in grasses are not common in monocots. Curr Genet 23: 160–165. doi:10.1007/BF00352016.
54. Wang R-J, Cheng C-L, Chang C-C, Wu C-L, Su T-M, et al. (2008) Dynamics and evolution of the inverted repeat-large single copy junctions in the chloroplast genomes of monocots. BMC Evol Biol 8: 1–14. doi:10.1186/1471-2148-8-36.
55. Terrab A, Paun O, Talavera S, Tremetsberger K, Arista M, et al. (2006) Genetic diversity and population structure in natural populations of Moroccan Atlas cedar (Cedrus atlantica; Pinaceae) determined with cpSSR markers. American Journal of Botany 93: 1274–1280. doi:10.3732/ajb.93.9.1274.
56. Grassi F, Labra M, Scienza A, Imazio S (2002) Chloroplast SSR markers to assess DNA diversity in wild and cultivated grapevines. Vitis.
57. Melotto-Passarin D, Tambarussi E, Dressano K, De Martin V, Carrer H (2011) Characterization of chloroplast DNA microsatellites from Saccharum spp and related species. Genet Mol Res 10: 2024–2033.
58. Kim K-J, Lee H-L (2004) Complete Chloroplast Genome Sequences from Korean Ginseng (Panax schinseng Nees) and Comparative Analysis of Sequence Evolution among 17 Vascular Plants. DNA Research 11: 247–261. doi:10.1093/ dnares/11.4.247.
59. Kim J-S, Jung J, Lee J-A, Park H-W, Oh K-H, et al. (2006) Complete sequence and organization of the cucumber (Cucumis sativus L. cv. Baekmibaekdadagi) chloroplast genome. Plant Cell Rep 25: 334–340. doi:10.1007/s00299-0050097-y.
60. Kuang D-Y, Wu H, Wang Y-L, Gao L-M, Zhang S-Z, et al. (2011) Complete chloroplast genome sequence of Magnolia kwangsiensis (Magnoliaceae): implication for DNA barcoding and population genetics. Genome 54: 663–673.
61. Kim Y-K, Park C, Kim K-J (2009) Complete chloroplast DNA sequence from a Korean endemic genus, Megaleranthis saniculifolia, and its evolutionary implications. Mol Cells 27: 365–381. doi:10.1007/s10059-009-0047-6.
62. Yi D-K, Kim K-J (2012) Complete Chloroplast Genome Sequences of Important Oilseed Crop Sesamum indicum L. PLoS ONE 7: e35872. doi:10.1371/journal.pone.0035872.
63. Raboin LM, Carreel F, Noyer J-L, Baurens F-C, Horry JP, et al. (2005) Diploid ancestors of triploid export banana cultivars: molecular identification of 2n restitution gamete donors and n gamete donors. Molecular Breeding 16: 333– 341.

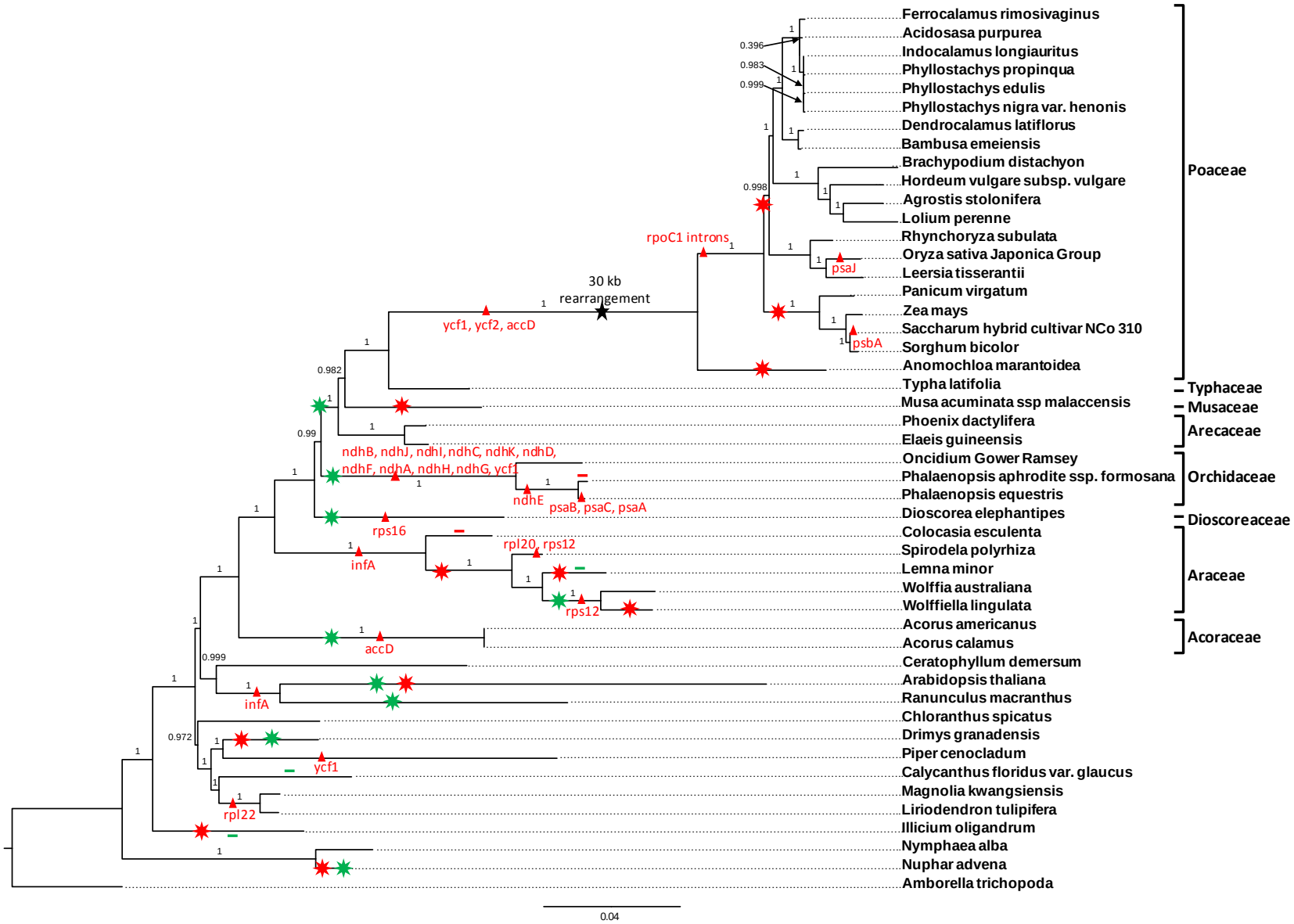


Figure S1 : Maximum likelihood phylogenetic analysis based on 79 chloroplast's protein coding genes of 46 basal angiosperms and monocotyledons and 3 Dicotyledons. The tree has a -lnL of -527912.066159. Support values for ML are provided at the nodes. Gene losses in chloroplast genomes are indicated with red triangles. Green and red stars represent partial or total IR gain of genes belonging respectively to LSC or SSC relative to *A. trichopoda* structure. Green and red minus signs represent loss of one of the two partial or complete gene copies belonging to IR respectively to become member of LSC or SSC relative to *A. trichopoda* structure.

Supplementary materials

This section contains 2 supplemental figures (S1 and S2) and three supplemental tables (S1, S2 and S3).

Figure S2 : Localization of cp DNA inserted in the nuclear genome of *M. acuminata*.

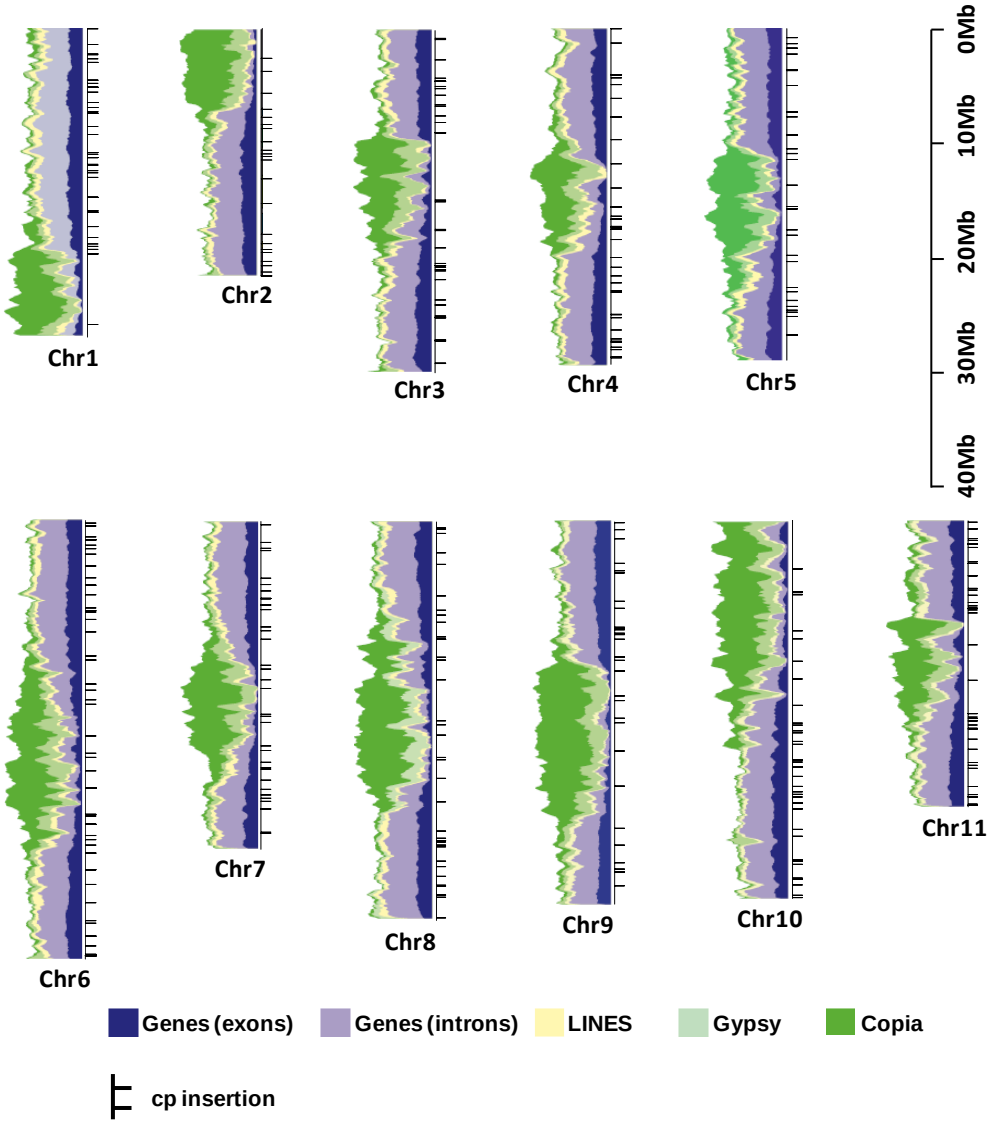


Table S1 : Chloroplast genomes compared with the *M. acuminata* chloroplast.

Taxa	Genome size	GenBank accession
Monocotyledons		
Ferrocalamus rimosivaginus	139,467	NC_015831
Acidosasa purpurea	139,697	NC_015820
Indocalamus longiauritus	139,668	NC_015803
Phyllostachys propinqua	139,704	NC_016699
Phyllostachys edulis	139,679	NC_015817
Phyllostachys nigra var. henonis	139,839	NC_015826
Dendrocalamus latiflorus	139,394	NC_013088
Bambusa emeiensis	139,493	NC_015830
Brachypodium distachyon	135,199	NC_011032
Hordeum vulgare ssp. vulgare	136,462	NC_008590
Agrostis stolonifera	136,584	NC_008591
Lolium perenne	135,282	NC_009950
Rhynchoryza subulata	136,303	NC_016718
Oryza sativa Japonica Group	134,525	NC_001320
Leersia tisserantii	136,551	NC_016677
Panicum virgatum	139,619	NC_015990
Zea mays	140,384	NC_001666
Saccharum hybrid cultivar NCo 310	141,182	NC_006084
Sorghum bicolor	140,754	NC_008602
Anomochloa marantoidea	138,412	NC_014062
Typha latifolia	161,572	NC_013823
Phoenix dactylifera	158,462	NC_013991
Elaeis guineensis	156,973	NC_017602
Oncidium Gower Ramsey	146,484	NC_014056
Phalaenopsis aphrodite ssp. Formosana	148,964	NC_007499
Phalaenopsis equestris	148,959	NC_017609
Dioscorea elephantipes	152,609	NC_009601
Colocasia esculenta	162,424	NC_016753
Spirodela polyrhiza	168,788	NC_015891
Lemna minor	165,955	NC_010109
Wolffia australiana	168,704	NC_015899
Wolffiella lingulata	169,337	NC_015894
Acorus americanus	153,819	NC_010093
Acorus calamus	153,821	NC_007407
Dicotyledons		
Ceratophyllum demersum	156,252	NC_009962
Arabidopsis thaliana	154,478	NC_000932
Ranunculus macranthus	155,129	NC_008796
Basal Angiosperms		
Chloranthus spicatus	157,772	NC_009598
Drimys granadensis	160,604	NC_008456
Piper cenocladum	160,624	NC_008457
Calycanthus floridus var. glaucus	153,337	NC_004993
Magnolia kwangsiensis	159,667	NC_015892
Liriodendron tulipifera	159,886	NC_008326
Illicium oligandrum	148,553	NC_009600
Nymphaea alba	159,930	NC_006050
Nuphar advena	160,866	NC_008788
Amborella trichopoda	162,686	NC_005086

Table S2: Distribution of simple sequence repeats (SSRs) loci in the *M. acuminata* chloroplast genome.

Region location	Unit length	SSR	Size	Start	End
LSC	5	(CTTAT)3	15	25	39
LSC	5	(AATAA)3	15	4300	4314
LSC	4	(ATT)4	12	5432	5443
LSC	1	(A)15	15	5455	5469
LSC	4	(ATTG)3	12	5893	5904
LSC	1	(C)10	10	8004	8013
LSC	4	(TATT)5	20	16007	16026
LSC	3	(TAT)4	12	16049	16060
LSC	3	(TAT)4	12	16061	16072
LSC	5	(TATTA)3	15	16073	16087
LSC	3	(TAA)5	15	16584	16598
LSC	1	(T)12	12	20615	20626
LSC	2	(AT)5	10	21986	21995
LSC	4	(AATA)3	12	30171	30182
LSC	2	(TA)6	12	31687	31698
LSC	5	(TTTAT)3	15	33492	33506
LSC	1	(A)10	10	34538	34547
LSC	4	(TTTG)3	12	35573	35584
LSC	1	(T)11	11	40036	40046
LSC	5	(TATAG)3	15	49486	49500
LSC	4	(CAAA)3	12	50026	50037
LSC	2	(AT)9	18	51235	51252
LSC	2	(TA)7	14	51297	51310
LSC	3	(AAT)4	12	54057	54068
LSC	1	(T)11	11	58405	58415
LSC	1	(T)12	12	58766	58777
LSC	1	(A)10	10	60414	60423
LSC	2	(TA)5	10	60489	60498
LSC	4	(ATAA)3	12	63130	63141
LSC	1	(A)10	10	64943	64952
LSC	4	(GAAA)3	12	67045	67056
LSC	4	(TCTT)3	12	68079	68090
LSC	3	(TTC)4	12	71867	71878
LSC	2	(TA)6	12	74114	74125
LSC	4	(AAGA)3	12	74683	74694
LSC	1	(T)12	12	75118	75129
LSC	1	(T)14	14	75548	75561
LSC	4	(AAAT)3	12	76130	76141
LSC	1	(T)13	13	85019	85031
IR	2	(GA)5	10	94674	94683
IR	3	(AGA)10	30	116275	116304
IR	3	(AGA)4	12	116527	116538
IR	6	(AAGCAG)5	30	117051	117080
IR	1	(T)10	10	117236	117245
IR	1	(A)12	12	117383	117394
IR	1	(A)10	10	117875	117884
IR	1	(A)15	15	120011	120025
IR	2	(AG)5	10	121964	121973
SSC	2	(TA)5	10	126873	126882
SSC	2	(AT)5	10	128655	128664
SSC	4	(AATA)3	12	129058	129069
SSC	4	(TTTA)3	12	131797	131808
SSC	1	(A)10	10	132407	132416
SSC	4	(AATA)3	12	134345	134356

Table S3 : Makers, associated primer, and expected length tested for the polymorphism analysis.

Markers	Forward	Reverse	length
mMaClRcp01*	CCCCCAGAAACGTATAGGAG	TTTCCCTTCGAATTCTGTCA	234
mMaClRcp02*	TAACCTCCCCAACCCTTCTT	GTGAATCCATGGAGGGTCAT	222
mMaClRcp03	CCTGCAAGTAAAGGGGGTTC	TTTCCGTGAGTCCTCGAAAT	197
mMaClRcp05	TGTTTTGTTCTTCCGCCAAT	TTGAACCCGGAATTGTCCTA	200
mMaClRcp06	TCGAGATAGATCGGGCGTTA	CGGGAGCATAATCTCACTTTG	270
mMaClRcp07	TCTCACTTTGTCTTGGGTTCTC	TCCCTAGACATTTTACCCCATA	277
mMaClRcp08	AAAAAGATTGGGCCGATTG	CCAATCCCAAGGATCCATAA	201
mMaClRcp09	TGAAATGATAATTGCAACGAAA	GGGCACAACTGGTACATTCC	198
mMaClRcp10	TGGCCAAGGGTAAAGATGTC	TCCGACCAAAATAGGCTTTG	391
mMaClRcp12	TTATCGGCTGTCTTGCGTTT	TACCGGGGATTTCGTGTAAA	211
mMaClRcp13	CATCCAGCAGGAATTGAACC	CTCAATTCGACGATCCAGAA	462
mMaClRcp14	GATTGATTGGTCCGAGGCTA	CCGGTCTTTGGGAAGTATCA	299
mMaClRcp15	CCTTGTTCATTCGCAACAAA	TCGAATTGGAAAACGGAAAG	182
mMaClRcp16	GCTTGTGGGGTCAAATCAAT	CCAATCACCGTTCACCTTTT	216
mMaClRcp17	ACAAGTTCCCGGATGACAAG	GGATCCACTTTTTGGGGAAT	388
mMaClRcp18	TGGGTTTCTACCAATGAGCA	TTGCTCATTCTCATCGTTGC	165
mMaClRcp19*	GGACCGTATCGTGGAACAAT	GCGGATTCTTTTCATGTTCA	226
mMaClRcp20*	CGAAACGGGTGGTGATCTAT	GGGGAATGAACATTTGTTTGA	217
mMaClRcp21	AACGGAACTCCCTTTTGGTT	TTTGAATGGTCGATTCGTGA	380
mMaClRcp22	GATTGGATGGGAATGAATCG	CCCTCTTTTTCTTGGTTGGA	467
mMaClRcp23	TGCACCAGAACAACTGGAAA	CATCATTCGCATACCTGTGG	211
mMaClRcp24	TGAAAATTCTTTTGTTTCTTATATGC	TTTTCTAACGATTTCGACACCT	168
mMaClRcp25*	AATAACGGGACCAAAACC	TCCTTCCTTCCATTCTCA	287
mMaClRcp26	TTTCTGTTTCCGGTGGTA	ATCTTACCCGGATCTTCG	223
mMaClRcp27*	CGGTTCAGGGTACGAATA	CCCCAAAAGTAAAAAGTGG	207
mMaClRcp28	AACGTCGATGGAGACGTA	CATTTGATTCGTCGATCC	127
mMaClRcp29*	AGTTGGTACCACCCAACC	GGCGGAAATCCAATATCT	283
mMaClRcp30*	AACAAACATTGGGTTTGG	AGTCCCTCCCTACAACTCA	279
mMaClRcp31*	TCAACGAATGAAGCAGGT	TATATGCGTTTCCGGGTA	292
mMaClRcp32*	ACCCCCGACACATAAAAT	CCGCTTCTATGGGATCTT	275
mMaClRcp33*	GGATGCATACGGTTCAAA	AAAGGCCCATTCAGAAAC	262
mMaClRcp34*	TGGTGCCTCCTAATTTTG	CGGGAATTGAGACAGTTG	250

*Highly polymorphic markers tested for the splitting of chloroplastic group II in *M. acuminata* subspecies.

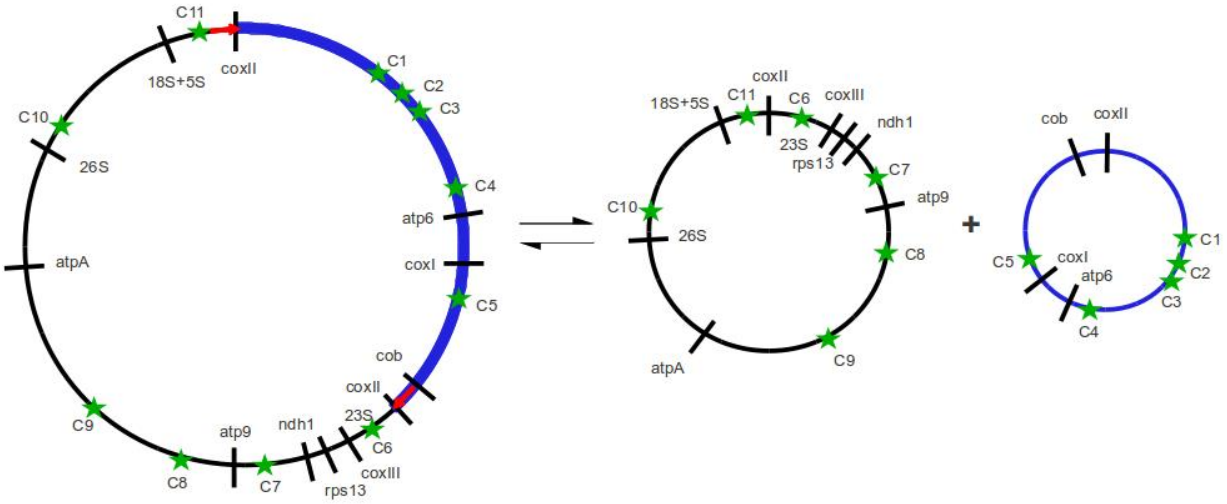


Figure 16 : Structure tricirculaire du génome mitochondrial de *Brassica campestris*. La recombinaison entre les deux copies répétées de 2 kb (contenant le gène coxII ; flèches rouges) réorganise le cercle principal de 218 kb en deux cercles, l'un de 135 kb (noir) et l'autre de 83 kb (bleu). Les traits noirs indiquent les gènes mitochondriaux, les étoiles vertes représentent les gènes chloroplastiques (D'après Palmer, 1992)

I.3 Assemblage partiel du génome mitochondrial du bananier

Le génome mitochondrial des angiospermes est de taille très variable, la majorité se trouvant entre 300 et 600 kb avec un minimum de 200 kb chez certaines Malvacées (Palmer, 1992) et un maximum pouvant dépasser les 1.5 Mb chez certaines Cucurbitacées (Alverson et al., 2011). Le génome mitochondrial des plantes se compose d'un set minimal de 26 gènes codant des protéines dont la majorité code pour des sous-unités des complexes de la chaîne respiratoire (Knoop, 2004). Bien que l'évolution des gènes mitochondriaux des angiospermes soit plus lente que celle des gènes nucléaires et chloroplastiques (Palmer and Herbon, 1988; Zhu et al., 2007), des variations importantes de la structure globale de la mitochondrie ainsi que de l'ordre des gènes ont été observées (Palmer and Herbon, 1988).

Le génome mitochondrial des plantes présente une structuration multipartite organisée en un ou plusieurs cercles d'ADN (D'Hont et al., 1987; Palmer, 1992). L'origine de ces différents cercles serait la présence de régions répétées qui, par des mécanismes de recombinaisons, entraîneraient la structuration du génome en plusieurs cercles d'ADN (**Figure 16**).

Une approche identique à celle réalisée pour l'assemblage du génome chloroplastique du bananier a été testée pour assembler le génome mitochondrial. Lors de cette tentative, le génome utilisé comme référence a été celui de *Phoenix dactylifera* (Fang et al., 2012). Cependant, la présence de séquences chloroplastiques (Fang et al., 2012) connues pour s'intégrer de manière récurrente dans le génome mitochondrial (Notsu et al., 2002; Palmer, 1992) a conduit à un envahissement de l'assemblage mitochondrial par le génome chloroplastique complet lors des étapes itératives. Cet envahissement est la conséquence de la combinaison de deux phénomènes : la récupération par recherche de similarité des séquences chloroplastiques présentes dans le génome mitochondrial, et un rapport déséquilibré en quantité de séquence en faveur du chloroplaste. Dans ces conditions, chaque fois qu'une séquence de chloroplaste est présente dans un contig mitochondrial, l'élongation consensus "penche" vers le chloroplaste et reconstruit le génome chloroplastique complet. L'utilisation de paramètres d'alignement et d'assemblage plus exigeants a permis de résoudre en partie ces problèmes d'envahissement du génome mais résulte en un assemblage très partiel. Cette approche n'a donc pas permis d'obtenir un assemblage satisfaisant du génome mitochondrial de bananier.

Par contre, lors des recherches d'hétérozygotie structurale chez 'Pahang', décrite dans le chapitre 2, plusieurs scaffolds de grande taille non ancrés aux chromosomes ont révélé des

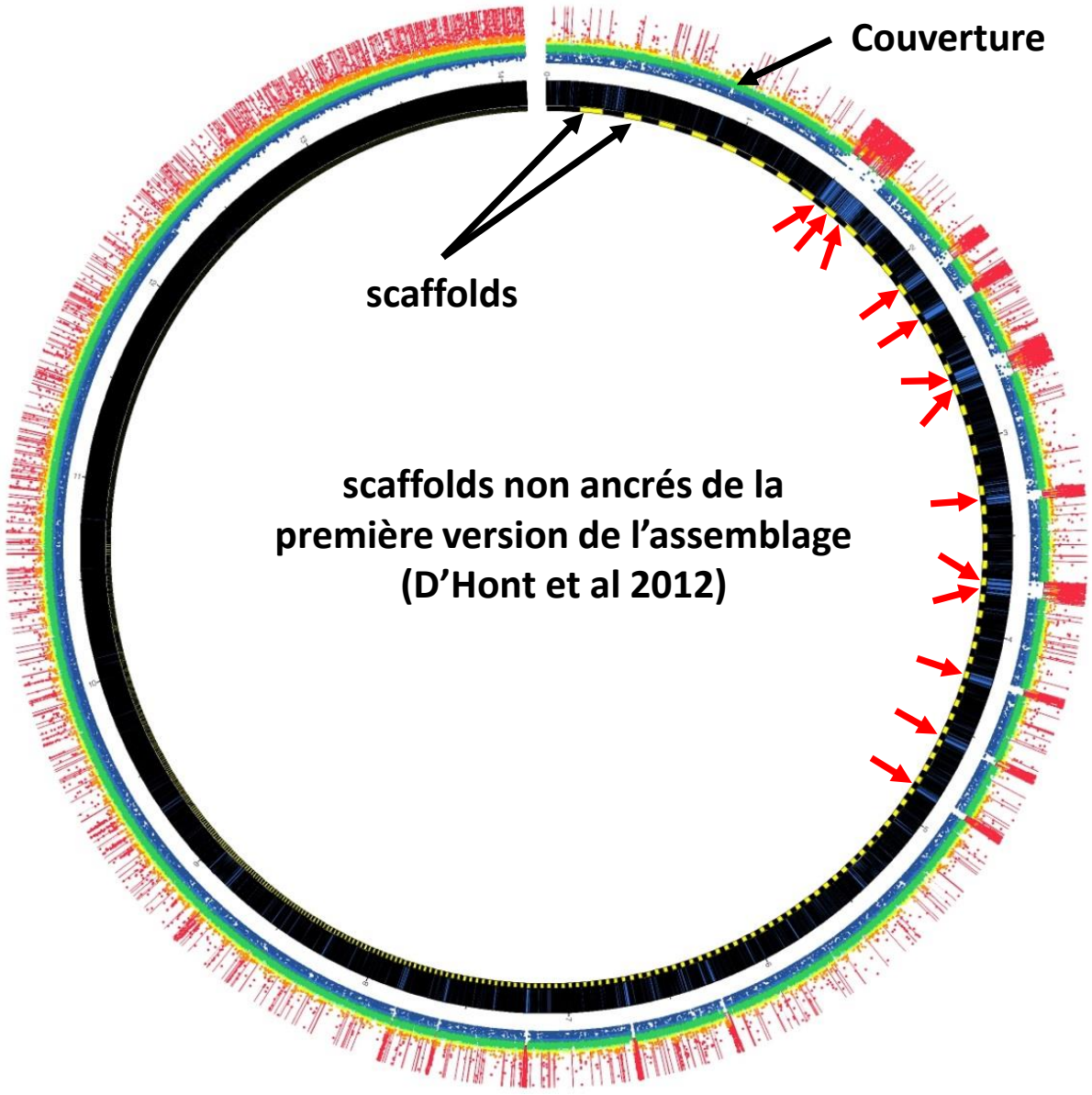


Figure 17 : Représentation Circos simplifiée des couvertures de la banque de 5 kb obtenue par re-séquençage de l'accession 'Pahang'. La couverture moyenne des lectures pairées est calculée sur une fenêtre de 1 kb. La couverture moyenne est codée en utilisant le code couleur suivant : quand la couverture est supérieure à quatre fois la couverture attendue, une barre rouge est dessinée. Pour toute autre valeur un point coloré est dessiné. Les fenêtres présentant une couverture supérieure à 5/4, 4/3, 3/2 et 2 fois la couverture attendue sont représentées par un point vert clair, jaune, orange et rouge respectivement. Les fenêtres présentant une couverture inférieure à 3/4, 2/3 et 1/2 fois la couverture médiane attendue sont représentées avec des points verts foncés, bleus et bleus foncés respectivement. Les fenêtres présentant une couverture comprise entre 3/4 et 5/4 sont représentées par des points verts. L'étude des couvertures de la banque 5 kb générée sur 'Pahang' permet d'observer de grands scaffolds (blocks noirs et jaunes) présentant une couverture anormalement élevée (flèches rouge).

couvertures anormalement élevées par rapport aux couvertures attendues (**Figures 17**). La recherche de similarités entre la séquence de ces scaffolds et la base de données nucléotidiques ('Nucleotidique collection') du National Center for Biotechnology Information (NCBI) a révélé comme seule similarité significative des gènes mitochondriaux. Ceci indique clairement l'appartenance de ces scaffolds au génome mitochondrial.

Finalement, les scaffolds mitochondriaux ont été recherchés et identifiés dans la nouvelle version de l'assemblage du bananier par similarité (BLAST) avec les 37 gènes mitochondriaux codant des protéines de *Phoenix dactylifera* (Fang et al., 2012). Tous les gènes mitochondriaux de *Phoenix dactylifera* ont pu être alignés sur des scaffolds de l'assemblage. L'appartenance de ces scaffolds au génome mitochondrial de la banane a été validée par deux tests:

1. Vérification par approche de similarité réciproque. Chaque scaffold identifié comme contenant un gène mitochondrial a été utilisé pour une recherche de similarité dans la base de données nucléotidiques complète du NCBI; le meilleur hit en dehors du genre *Musa* a été analysé et doit correspondre à une séquence mitochondriale.
2. Vérification que la couverture des lectures de 'Pahang-HD' sur ces scaffolds est homogène et plus élevée que l'attendu pour un scaffold 'nucléaire'. Cette couverture, plus importante, est attendue et s'explique par la présence de plusieurs mitochondries par cellule contre un seul noyau. Dans notre cas, le ratio se situe aux alentours de quatre copies du génome mitochondrial pour une copie de génome nucléaire.

Tous les scaffolds dans lesquels des gènes mitochondriaux ont été identifiés ont été ainsi validés et assignés au génome mitochondrial. Ils ont ainsi pu être retirés de la partie non ancrée de la séquence de référence du génome nucléaire.

Finalement, 12 scaffolds mitochondriaux ont été identifiés. Leur taille cumulée est supérieure à 7.2 Mb avec une N50 de 616 kb et une taille allant de 172 kb à 3.77 Mb. L'ensemble des 12 scaffolds comprend 37.5kb (0.52%) de N.

Chez les angiospermes, aucun génome mitochondrial ne présente une taille aussi importante. Le génome mitochondrial du bananier serait environ quatre fois plus gros que le plus grand génome mitochondrial rapporté chez les angiospermes (*Cucumis sativus*, 1 685 kb). Des tailles similaires ont été observées chez des gymnospermes, *Picea abies* et *Picea glauca* avec des tailles respectives de 5.5 et 6 Mb (Jackman et al., 2014).

Pour finaliser cet assemblage, nous avons tenté d'isoler les mitochondries du bananier mais sans succès jusqu'à présent.
Un total de 24 marqueurs SSR a été développé sur la base de ces scaffolds mitochondriaux et a été testé sur cinq bananiers appartenant à l'espèce *Musa acuminata* et une accession des espèces *Musa balbisiana* et *Musa boman*. Parmi ces 24 marqueurs, 7 se sont révélés polymorphes intra *Musa acuminata* et 12 ont été identifiés polymorphes inter-espèces.

Chez le bananier, la transmission des génomes cytoplasmiques est particulière puisque, contrairement à la majorité des angiospermes où les deux organites (chloroplaste et mitochondrie) sont à hérédité maternelle, le chloroplaste est à hérédité maternelle alors que la mitochondrie est à hérédité paternelle (Fauré et al., 1994). Ces particularités font des marqueurs dérivés de ces organites un outil particulièrement intéressant en complément des marqueurs nucléaires, pour mieux comprendre l'origine des bananiers cultivés et pour vérifier et valider les croisements dans le cadre d'analyses génétiques ou de programmes d'amélioration.

Conclusion

A l'issue de ce travail, nous avons maintenant une meilleure connaissance des différents compartiments génomiques du bananier avec un draft de génome mitochondrial (sous la forme de douze scaffolds), un génome chloroplastique complet et une séquence de référence du génome nucléaire grandement améliorée. L'amélioration de cette séquence de référence du génome du bananier aura une répercussion importante sur la qualité des analyses et interprétations qui seront réalisées par la communauté scientifique sur la base de cette référence. Il sera, par ailleurs, intéressant d'analyser plus en détail le génome mitochondrial (structure, composition en séquences) qui semble avoir une structure particulière chez la banane.

Chapitre II : Développement d'outils pour la détection de réarrangements chromosomiques : application à l'analyse des accessions 'Pahang' et 'PKW'

La disponibilité récente de séquences de référence du génome pour de nombreuses espèces a permis un développement important des recherches sur les petites et moyennes variations structurales entre génomes de différentes espèces (Carbone et al., 2006; Girirajan et al., 2009; Newman et al., 2005) ou au sein d'une même espèce (Bashir et al., 2010; Carrier, 2011; Feulner et al., 2013; Hormozdiari et al., 2011; Kim et al., 2008; Korbel et al., 2007; McKernan et al., 2009; Quinlan et al., 2010; Sabot et al., 2011; The 1000 Genomes Project Consortium, 2012; Tuzun et al., 2005; Xing et al., 2009). Ces recherches ont conduit au développement de différents types d'outils informatiques présentés en introduction. Une trentaine de programmes ont été développés pour identifier les variations structurales sur la base de données de re-séquençage (présenté en introduction). Il est cependant important de noter que la plupart de ces algorithmes ont été développés pour le génome humain, génome pour lequel on possède une référence très complète et de très bonne qualité. Ces algorithmes sont utilisés en particulier dans le cadre des recherches sur les variations structurales impliquées dans le développement des cancers (Bashir et al., 2008; Hampton et al., 2009; Northcott et al., 2012). Une part importante de ces programmes a pour but d'identifier des variations de structures de type insertion ou délétion et se base sur l'inspection des distributions des couvertures. Ces programmes ne sont efficaces que si la variation recherchée est de taille relativement modeste et/ou le génome de référence est très proche du génome re-séquencé. Ces approches ne semblent donc pas appropriées pour la recherche de grandes variations structurales comme celles prédites chez le bananier, d'autant plus que ces programmes ne permettent pas de détecter les translocations qui représentent une grosse partie des variations supposées chez le bananier. D'autres approches se basant sur l'identification de lectures discordantes semblent plus adaptées à notre problématique puisqu'elles permettent d'identifier un plus grand nombre de types de variations structurales comme les translocations, inversions, délétions et insertions. Cependant, les programmes développés sont adaptés à l'analyse de génomes re-séquencés très proches de la référence, facilitant ainsi l'alignement des lectures obtenues. Ainsi, l'utilisation de ces approches avec des génomes re-

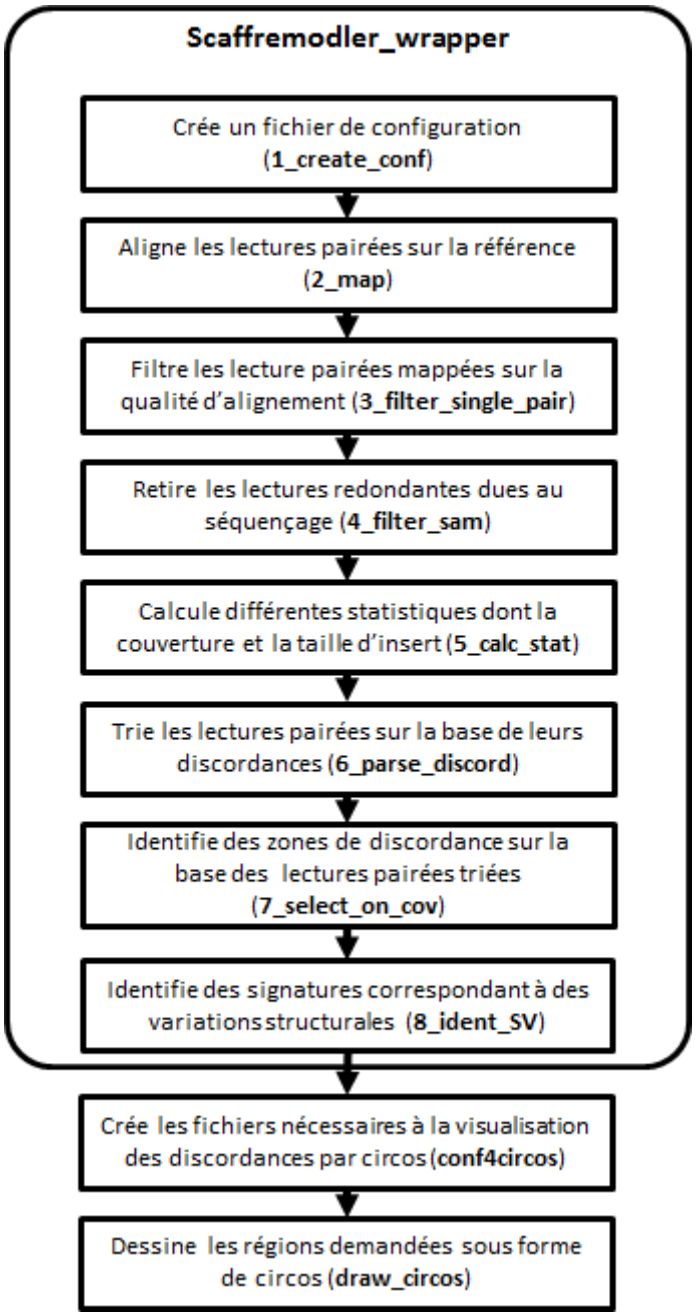


Figure 18 : Schéma décrivant l'enchaînement des programmes développés, pour rechercher les variations structurales dans un génome re-séquencé par rapport à une séquence de référence.

séquencés relativement divergents de la séquence de référence (dans notre cas entre sous-espèces et espèces) entraîne la détection d'un grand nombre d'évènements qui ralentit considérablement l'analyse et noie le signal dans un bruit de fond qui peut empêcher la détection des variations. Ces considérations sont au moins en partie responsables des problèmes (temps de calcul interminable, erreurs de mémoire et nombre énorme de variations détectées) rencontrés lorsque nous avons testé certains de ces outils (SV detect, Breakdancer, VariationHunter) sur notre jeu de données. Par ailleurs, chez le bananier les évènements structuraux recherchés sont des variations structurales de grande taille (plusieurs centaines de milliers de bases) et des événements rares (pas plus de quelques translocations supposées entre les différents groupes). Pour toutes ces raisons, les approches classiques de détection de variations structurales ne nous ont pas semblé adaptées à notre problématique.

J'ai donc développé un pipeline bioinformatique pour la recherche de réarrangements chromosomiques. Ce pipeline a ensuite été utilisé par analyser les accessions 'Pahang' et 'PKW'.

II.1 Développement et test d'un pipeline bioinformatique pour la détection de réarrangements chromosomiques

II.1.1 Le pipeline

Nous avons mis en place une stratégie de détection des variations structurales adaptée aux variations de grande taille et entre des génomes divergents. Cette stratégie est basée sur l'analyse de banques (ensemble de lectures pairées issues du séquençage des extrémités d'un grand nombre de séquences ADN) de grande taille analysées au moyen d'un pipeline informatique de 10 programmes séquentiels que j'ai écrits en langage de programmation Python. Neuf de ces programmes ont également été utilisés pour l'amélioration de la séquence de référence décrite dans la publication 1 de cette thèse (cf chapitre I). Ces programmes sont disponibles en ligne de commande et ont également été adaptés à la plateforme GALAXY, un des systèmes graphiques principaux pour exécuter des workflows (=flux de travaux) (Schatz et al., 2012). Ce pipeline peut être décrit succinctement comme suit (**Figure 18**):

(i) Le premier programme (*1_create_conf*) collecte des informations sur la séquence de référence utilisée et crée un fichier de configuration qui permet de ne renseigner les différentes options qu'une seule fois au cours de l'analyse. Cependant, cette étape n'est pas

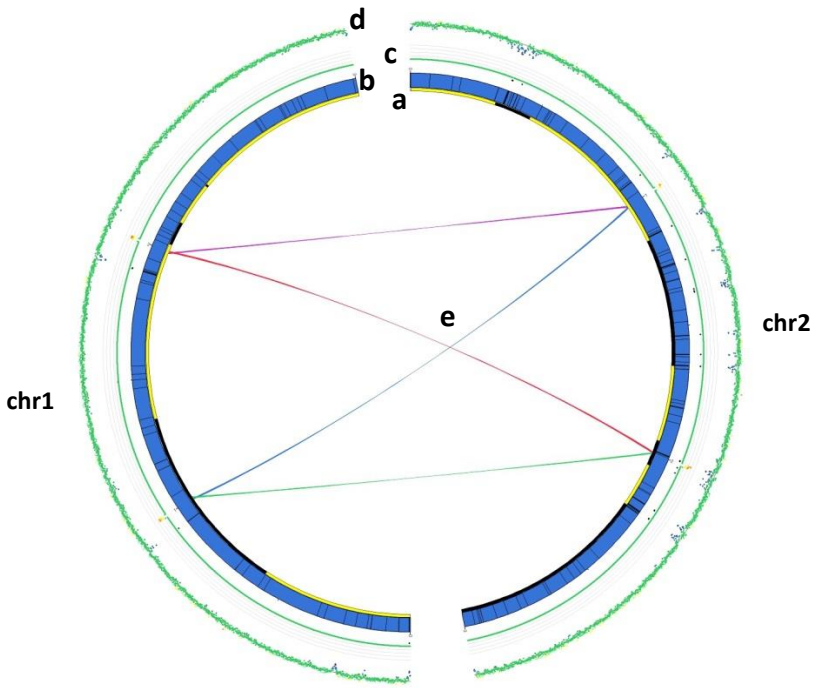


Figure 19 : Exemple de représentation graphique générée en sortie du pipeline de détection des variations structurales. Du cercle le plus central vers le plus extérieur : (a) localisation des scaffolds dans le génome de référence. Les scaffolds sont identifés par une alternace de blocs de couleur jaune et noire. Le second cercle (b) représente le génome de référence en bleu et les régions noires indiquent les séquences inconnues (les régions contenant des N). Le troisième cercle (c) représente la proportion en lectures discordantes s'alignant sur une fenêtre de 1 kb. Cette proportion est codée en utilisant le code couleur suivant : noire quand il n'y a pas de lecture s'alignant dans la fenêtre, vert, jaune, et orange pour les fenêtres qui contiennent moins de 25%, 50% et 75% respectivement. Les fenêtres qui contiennent plus de 75% de lectures discordantes sont représentées par un point rouge. Le dernier cercle (d) représente la couverture moyenne des lectures pairées calculée sur une fenêtre de 1 kb. La couverture moyenne est codée en utilisant le code couleur suivant : quand la couverture est supérieure à quatre fois la couverture attendue, une barre rouge est dessinée. Pour toute autre valeur un point coloré est dessiné. Les fenêtres présentant une couverture supérieure à 5/4, 4/3, 3/2 et 2 fois la couverture attendue sont représentées par un point vert clair, jaune, orange et rouge respectivement. Les fenêtres présentant une couverture inférieure à 3/4, 2/3 et 1/2 fois la couverture médiane attendue sont représentées avec des points verts foncés, bleus et bleus foncés respectivement. Les fenêtres présentant une couverture comprise entre 3/4 et 5/4 sont représentées par des points verts. (e) Les régions chromosomiques non contiguës liées par des lectures pairées sont représentées par des liens colorés en fonction de l'orientation des lectures pairées s'alignant dans ces régions et de l'orientation attendue des lectures. En partant du postulat que l'orientation attendue est reverse-forward (rf), si les lectures ont une orientation reverse-forward (rf) mais que la distance entre les deux lectures sœurs est supérieure à la distance attendue, le lien dessiné sera rouge. Si cette distance est inférieure à celle attendue, le lien dessiné sera noir. Si les lectures ont une configuration d'alignement forward-reverse (fr) le lien dessiné sera bleu. Si les lectures ont une configuration d'alignement reverse-reverse (rr) ou forward-forward (ff) le lien dessiné sera respectivement violet ou vert.

nécessaire et les neuf programmes suivants peuvent être utilisés sans ce fichier de configuration.

(ii) Le second programme (*2_map*) aligne les lectures pairées sur la séquence de référence en utilisant au choix les logiciels Bowtie (Langmead et al., 2009), Bowtie2 (Langmead and Salzberg, 2012) ou BWA (Li and Durbin, 2010).

(iii) Le troisième programme (*3_filter_single_pair*) filtre les lectures pairées sur la base des informations contenues dans le fichier 'sam' généré par le deuxième programme. Le filtre peut être appliqué sur deux paramètres : la 'mapping quality' (MAPQ) et/ou des variables AS/XS. La MAPQ est une statistique calculée par le logiciel d'alignement qui est égale à $-10\log_{10}$(probabilité d'alignement à la mauvaise position). Les variables AS et XS sont calculées quand les logiciels Bowtie et Bowtie2 sont utilisés et correspondent pour une lecture au score d'alignement du meilleur alignement et au second meilleur alignement respectivement.

(iv) Le quatrième programme (*4_filter_sam*) est également une étape de filtre qui élimine les lectures dupliquées (identiques) et les paires de lectures qui n'ont pas été alignées.

(v) Le cinquième programme (*5_cal_stat*) calcule différentes statistiques par rapport à la taille de "l'insert" de la banque séquencée et la couverture de la séquence de référence.

(vi) Le sixième programme (*6_parse_discorde*) trie les lectures pairées sur la base de leur configuration d'alignement (orientation et taille "d'insert"). En effet, selon le mode de construction de la banque séquencée, les lectures pairées ont une orientation attendue. Si cette orientation attendue est reverse pour la première lecture et forward pour la seconde lecture (noté rf), les paires de lecture s'alignant différement, c'est à dire forward-reverse (fr), reverse-reverse (rr), forward-forward (ff) sont identifiées et triées dans différents fichiers. Les lectures présentant la bonne orientation mais ayant une distance les séparant supérieure (del) ou inférieure (ins) à un intervalle (attendu par rapport à la taille des segments séquencés) sont également triées dans différents fichiers.

(vii) Le septième programme (*7_select_on_cov*) identifie des paires de zones discordantes (zones non contiguës du génome liées par des lectures pairées) en se basant sur les lectures pairées triées par le sixième programme. Les principes et le fonctionnement de ce programme sont détaillés dans le point suivant (**II.1.2 Les programmes 7_select_on_cov et 8_ident_SV**).

(viii) Le huitième programme (*8_ident_SV*) recherche parmi les paires de zones discordantes identifiées par le septième programme des signatures correspondant à des variations

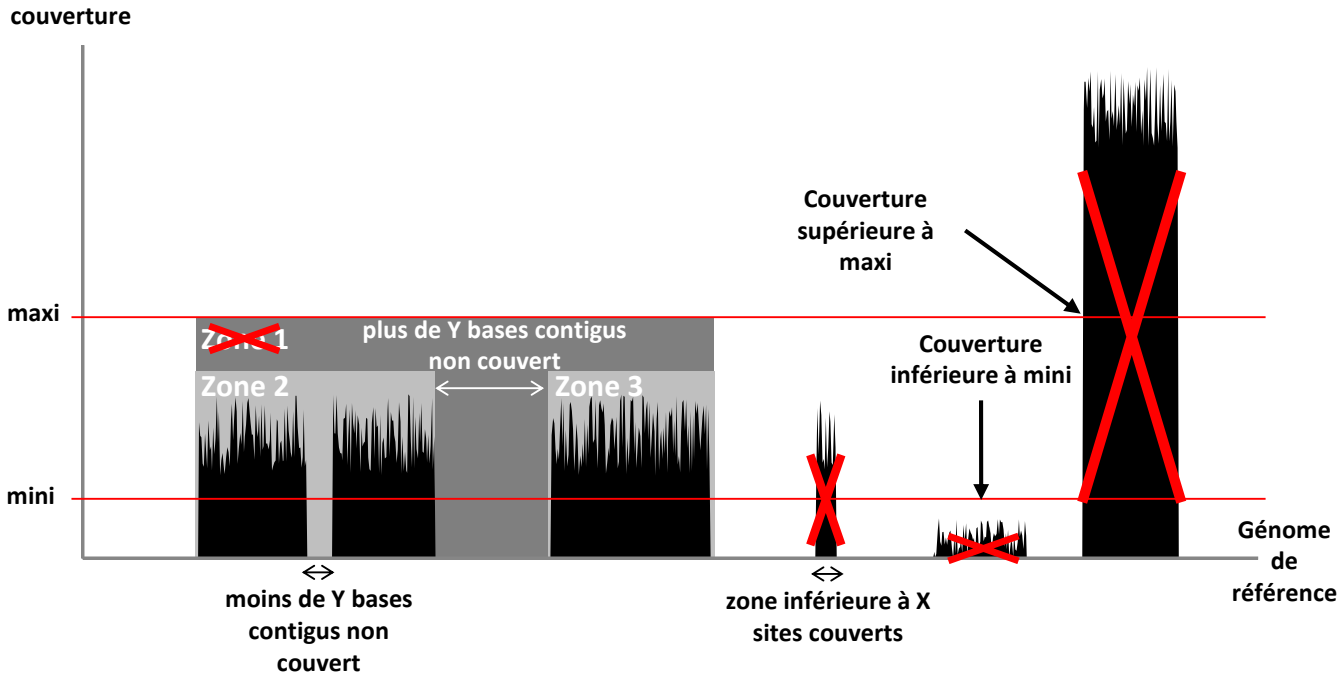


Figure 20 : Schématisation de l'identification des zones de discordance basée sur l'analyse des couvertures des lectures discordantes correspondant à la seconde étape du programme 7_select_on_cov.

structurales. Cette étape de recherche est développée plus en détail dans le point suivant (**II.1.2 Focus sur 7_select_on_cov et 8_ident_SV**).

(ix) Le neuvième programme (*conf4circos*) génère des fichiers nécessaires à la visualisation graphique des paires de zones discordantes identifiées par le septième programme.

(x) Le dixième programme (*draw_circos*) utilise les fichiers générés par le neuvième programme pour créer des représentations graphiques des paires de zones discordantes identifiées. Les informations contenues dans ces représentations graphiques sont décrites en **Figure 19**.

Les programmes 1 à 8 peuvent être utilisés séquentiellement mais ont également été englobés dans un programme (*Scaffremodler_wrapper*) qui permet de lancer la suite de traitement en une seule fois. Pour gagner du temps de calcul, une partie de ces programmes (*Scaffremodler_wrapper, 2_map* et *8_ident_SV*) est parallélisable.

II.1.2 Les programmes 7_select_on_cov et 8_ident_SV

Ces deux programmes constituent le coeur du pipeline puisque que ce sont ces programmes qui identifient les paires de zones discordantes (*7_select_on_cov*) et déduisent de l'analyse de ces paires de zones discordantes identifiées des variations structurales (*8_ident_SV*). Il m'a donc semblé important de décrire plus en détail leur principe et fonctionnement.

Le programme *7_select_on_cov* qui recherche des paires de zones discordantes travaille indépendamment sur chaque fichier contenant les lectures discordantes triées par le programme *6_parse_discorde* et s'exécute en cinq étapes :

(i) La couverture des lectures triées sur leur configuration d'alignement sur le génome de référence est calculée.

(ii) Des zones de discordance sont ensuite identifiées sur la couverture avec les critères suivants (**Figure 20**) : une zone est définie si elle présente au moins **X** sites couverts avec pas plus de **Y** sites contigus non couverts. La zone identifiée n'est pas conservée si la couverture médiane des sites couverts n'est pas comprise entre des valeurs de couvertures minimales (**mini**) et maximales (**maxi**). Ce seuil de couverture maximale permet d'éliminer en théorie les séquences répétées (très couvertes) qui risquent de créer un bruit important dans l'analyse. Le seuil de couverture minimale ainsi que la taille minimale de la zone permettent d'éliminer les faux positifs liés à des problèmes d'alignement des lectures.

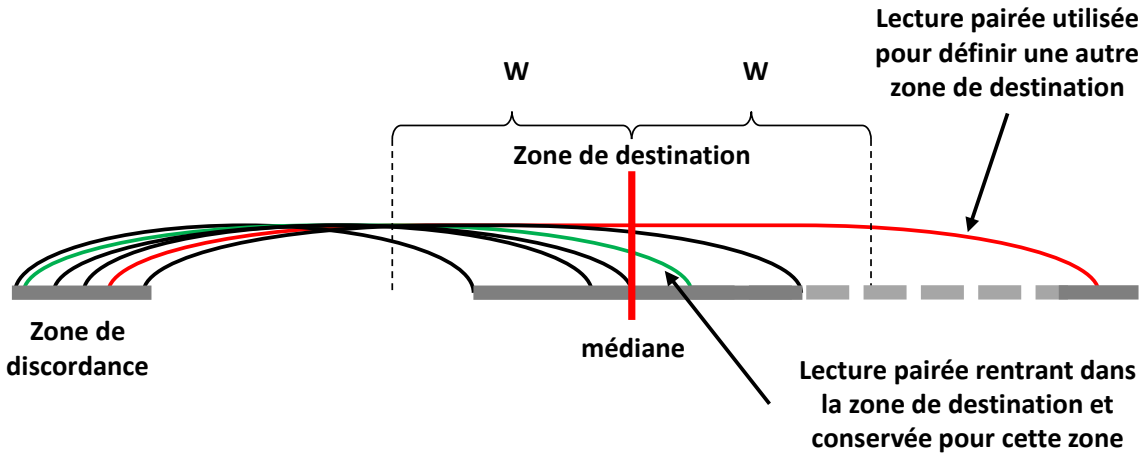


Figure 21 : Schématisation de l'identification de la zone de destination d'une zone de discordance correspondant à la troisième étape du programme 7_select_on_cov. Les lectures pairées (lien vert) dont une des sœurs entre dans la zone de destination déjà identifiée par les lectures pairées est conservée pour définir cette zone. La lecture pairée (lien rouge) dont une des sœurs s'aligne dans la zone de discordance mais la seconde sœur ne s'aligne pas dans la zone de destination sera utilisée pour rechercher une autre zone de destination. La valeur **W** est égale à la taille de la zone de discordance auquel est ajoutée une valeur égale à 3*(écart type de la taille d'insert).

(iii) Pour chacune des zones de discordance identifiées en (ii), les lectures sœurs des lectures s'alignant dans une zone de discordance sont utilisées pour identifier une ou des zones de destinations. Comme une zone de discordance peut avoir plusieurs zones de destinations, ces zones de destinations sont identifiées en suivant une approche itérative qui attribue les lectures de destination aux différentes zones possibles de destination **(Figure 21)**. Cette approche consiste en une première étape qui prend aléatoirement une des lectures discordantes s'alignant dans la zone de discordance et à attribuer comme première zone de destination la position de sa lecture sœur. L'étape d'itération consiste à ajouter successivement des lectures sœurs de celles s'alignant dans la zone de discordance afin de mieux définir les limites de cette zone de destination. La validation de l'appartenance de lectures sœurs à la zone de destination est effectuée si celles-ci s'alignent dans une zone comprise autour de la médiane des positions des lectures sœurs déjà attribuées à la zone, plus ou moins une valeur **W**. Cette valeur **W** est égale à la taille de la zone de discordance à laquelle est ajoutée une valeur **Z** = 3*(écart type de la taille "d'insert"). Si aucune lecture sœur supplémentaire ne peut être attribuée à la zone de destination, la recherche d'une nouvelle zone de destination est effectuée jusqu'à attribution de toutes les lectures sœurs des lectures s'alignant dans la zone discordante. Les limites des zones de destination et leurs zones de correspondance sont recalculées et filtrées sur les critères **X**, **mini** et **maxi** définis en (ii).

(iv) Les paires de zones identifiées de même type et adjacentes sont ensuite regroupées. Pour que deux paires de zones adjacentes soient regroupées, il faut qu'elles présentent le même type de discordance et que leurs zones de discordances et zones de destinations soient chacune suffisamment proches. La valeur de proximité est fournie par l'utilisateur sous forme d'un entier correspondant au nombre de bases maximum accepté entre les deux zones de discordances et les deux zones de destinations.

(v) Un score est ensuite appliqué aux paires de zones identifiées en se basant sur la moyenne des tailles d'une paire de zone de discordance et de destination et sur la couverture médiane par les lectures discordantes de ces zones. Le score est calculé de la façon suivante :

$$Score = 100 * f(couverture\ mediane) * g(taille\ de\ la\ zone)$$

$$f(x) = \begin{cases} \frac{1}{a} * x & (si\ x < a) \\ 1 & (si\ x \geq a) \end{cases}$$

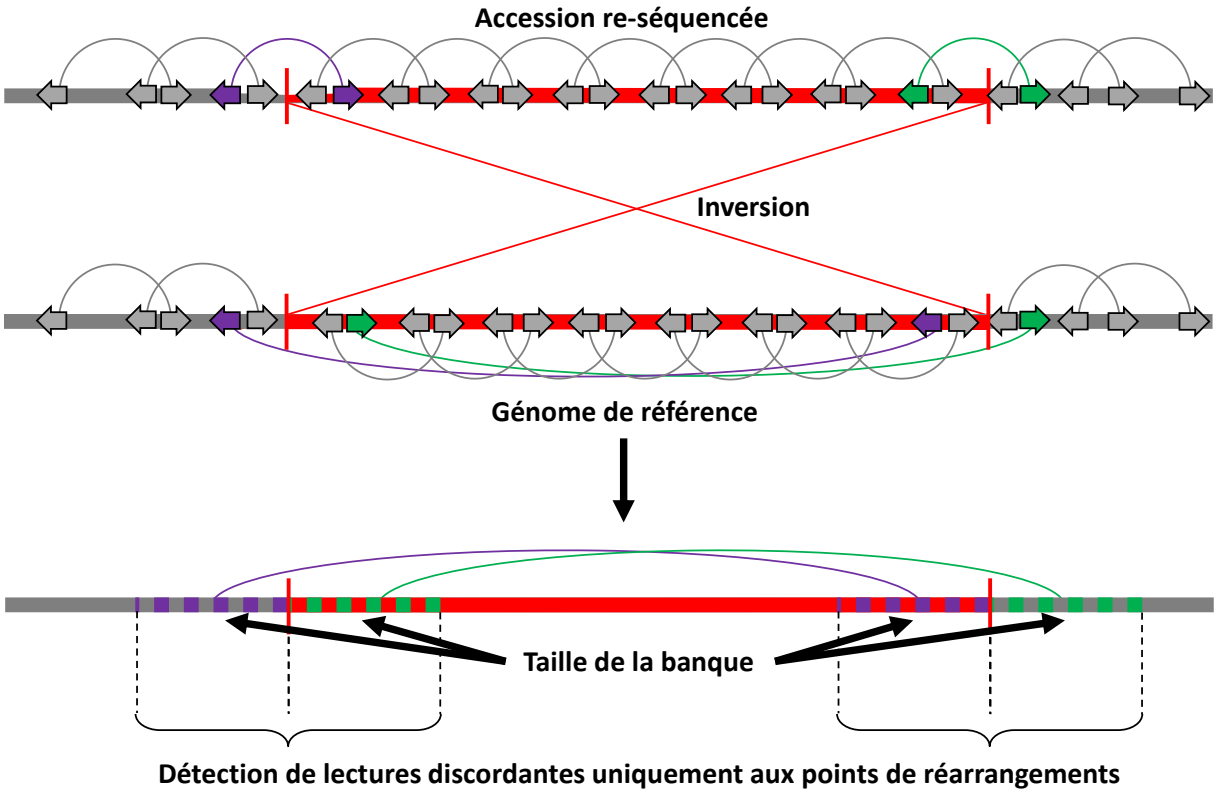


Figure 22 : Schématisation de la détection des bornes d'une inversion. Seules les lectures pairées (flèches) qui chevauchent le point de réarrangement (trait vertical rouge) permettent d'identifier l'inversion, ce qui conduit à l'identification de deux paires de zones discordantes. Ces zones font au maximum la taille maximale des séquences de la banque.

$$g(x) = \begin{cases} \frac{1}{b} * x & (si\ x < b) \\ 1 & (si\ x \geq b) \end{cases}$$

Avec **a**, la valeur minimale de couverture pour que f(x) soit égale à 1 et **b**, la valeur minimale de la taille d'une zone de discordance pour que g(x) soit égale à 1.
Le score obtenu pour chaque zone est compris entre 0 et 100. Seules les zones ayant un score supérieur à un seuil fixé par l'utilisateur sont conservées pour la suite du pipeline. Les paramètres **X, Y, mini, maxi, Z, a, b** sont des options du programme. Ces paramètres permettent de s'adapter au jeu de données à disposition (taille de "l'insert" attendue, couverture, divergence avec la séquence de référence, qualité de la séquence de référence, orientation attendue des lectures) en jouant sur la taille minimale de la zone de discordance détectée, la présence ou non d'interruptions dans la zone de discordance, le nombre minimal de lectures pour valider une zone et le score des zones détectées.

Le programme *8_ident_SV* recherche parmi les paires de zones discordantes identifiées la signature de variations structurales. Ce programme se base sur le principe que la recherche de variations structurales de grande taille en utilisant des données de re-séquençage, ne permet de détecter des lectures pairées discordantes qu'au niveau des points de réarrangements (**Figure 22**) et, plus précisément, sur une zone de la taille "d'insert" de la banque. Pour chaque type de variations structurales, on peut identifier des configurations de paires de zones discordantes typiques (signatures). Ce sont ces signatures que le programme *8_ident_SV* recherche. L'ensemble des signatures recherchées par le programme *8_ident_SV* et les figures Circos attendues sont présentées en **Annexe 1**. Ce programme recherche les déletions, les inversions, les duplications simples ou inversées et les translocations simples ou réciproques avec inversion ou non des régions transloquées. Les duplications et translocations sont recherchées à la fois au niveau intra- et inter-chromosomique. En cas de variations structurales situées à l'extrémité d'un chromosome, les signatures sont partielles et ne sont donc pas directement identifiées par ce programme *8_ident_SV*. Par contre, les lectures discordantes et les zones de discordance que ces variations engendrent sont visualisables sur les figures Circos générées par le programme *draw_circos*.

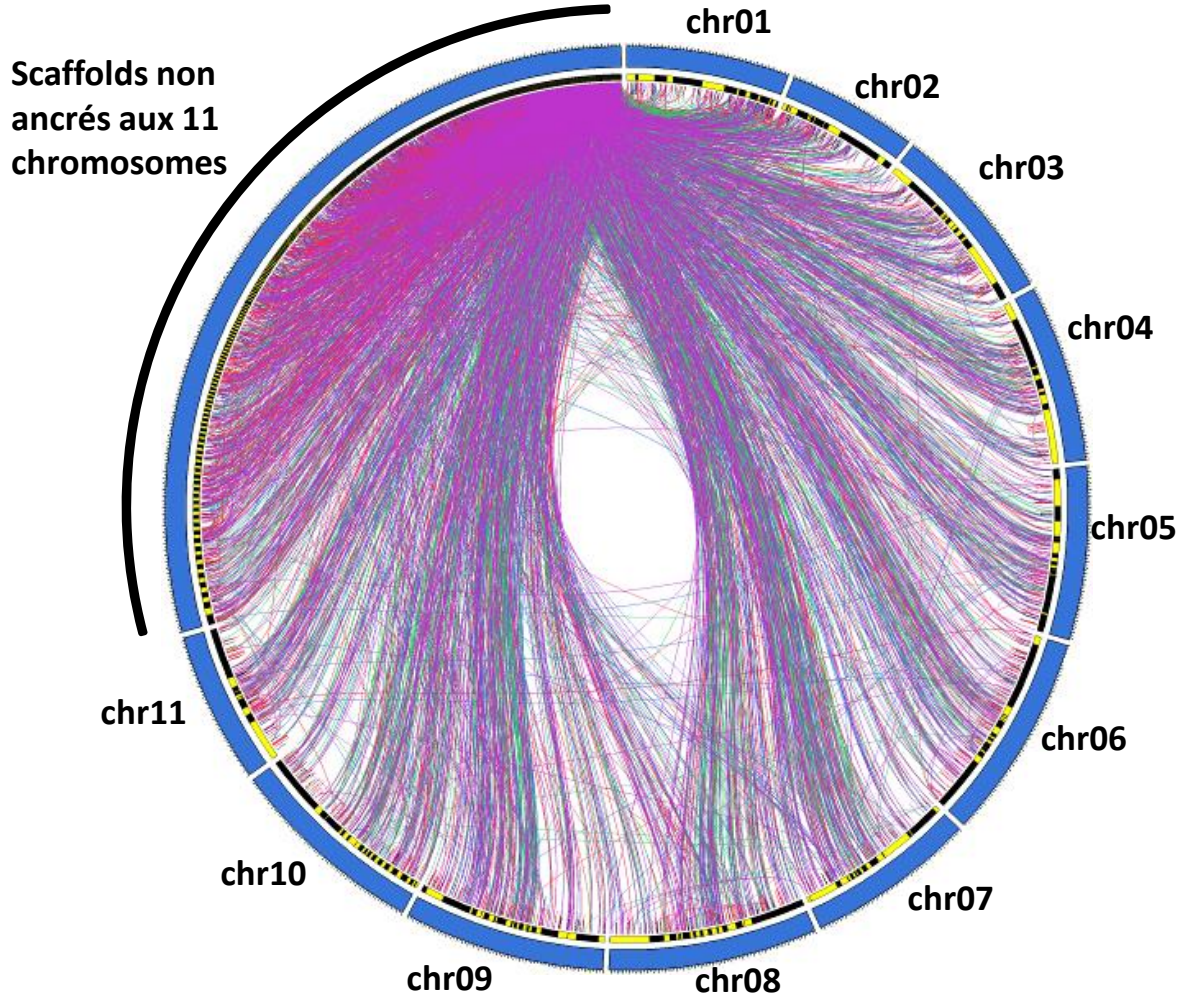


Figure 23 : Représentation Circos simplifiée des paires de zones discordantes identifiées par le pipeline de recherche de variations structurales avec la banque de 5 kb obtenue par re-séquençage de l'accession 'Pahang'.

II.2 Test du pipeline sur l'accession 'Pahang' en utilisant comme référence la version initiale de la séquence du génome A (*Musa acuminata*).

Ce pipeline a été testé sur l'accession 'Pahang', le parent de l'haploïde séquencé. En effet, lors de l'élaboration de la carte génétique de 'Pahang' qui a servi à l'ancrage de la séquence de référence du bananier, d'importantes distorsions de ségrégation ont été observées sur deux groupes de liaison correspondants aux chomosomes 1 et 4. Ces distorsions sont similaires à celles observés dans la descendance de 'Pisang Lilin' (Hippolyte et al., 2010) qui ont été interprété par les auteurs comme pouvant résulter d'une duplication à l'état hétérozygote d'une région du chromosome 1 dans une des deux versions du chromosome 4. L'accession 'Pahang' a donc été choisie pour tester notre pipeline et évaluer le potentiel de notre approche de re-séquençage pour détecter des variations structurales de grande taille.

Une banque de 5 kb, d'une couverture estimée de 60x, générée à partir de l'accession 'Pahang' a été analysée avec notre pipeline et en utilisant comme référence la version initiale de la séquence de référence du bananier (génome A, *Musa acuminata*). Les résultats ont révélé un très grand nombre de paires de zones discordantes pouvant correspondre à des variations structurales mais n'ont permis ni de détecter la duplication supposée entre les chromosomes 1 et 4 ni de montrer d'indice de la présence d'autres types de variations structurales simples impliquant ces deux chromosomes.

L'analyse des résultats a montré qu'un très grand nombre de paires de zones discordantes impliquait des scaffolds non ancrés aux 11 pseudo-molécules (ces scaffolds non ancrés représentaient 30% de l'assemblage initiale du génome de référence), ce qui créait un bruit très important et gênait fortement la détection de variations structurales (**Figure 23**).

D'autre part, la détection des variations structurales par ce pipeline est basée sur la reconnaissance de combinaisons de zones de discordances. La présence de 17% de N dans la version initiale de la séquence de référence du génome A gênait donc potentiellement fortement la détection de variations structurales en augmentant la probabilité de ne pas détecter l'un des liens signature par l'absence dans le génome de référence de la séquence correspondante.

Dans ce contexte, il est apparu nécessaire d'améliorer la séquence de référence du génome du bananier avant d'essayer d'identifier des variations structurales entre les bananiers. Ce travail d'amélioration de la séquence de référence du génome A du bananier a été décrit dans le chapitre I.

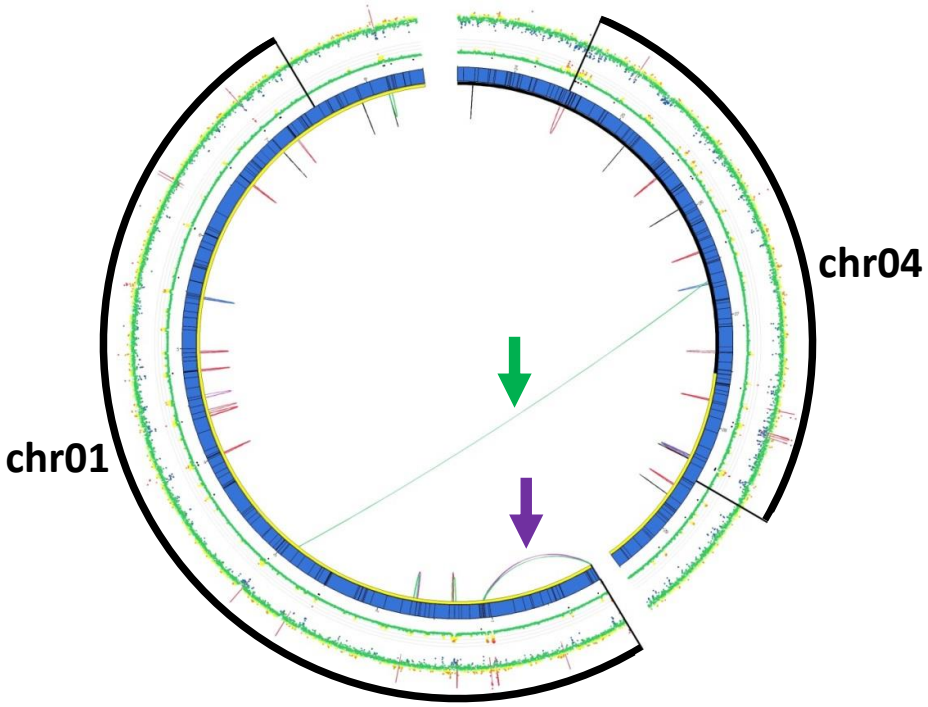


Figure 24 : Représentation Circos des paires de zones discordantes identifiées par alignement de la banque 5 kb de 'Pahang' sur les régions les plus distordues des chromosomes 1 et 4 du génome de référence (*Musa acuminata*). Les zones en noir représentent les régions les plus distordues des chromosomes 1 et 4. La flèche violette indique deux paires de zones discordantes à l'extrémité du chromosome 1. La flèche verte indique une paire de zones discordantes reliant les chromosomes 1 et 4.

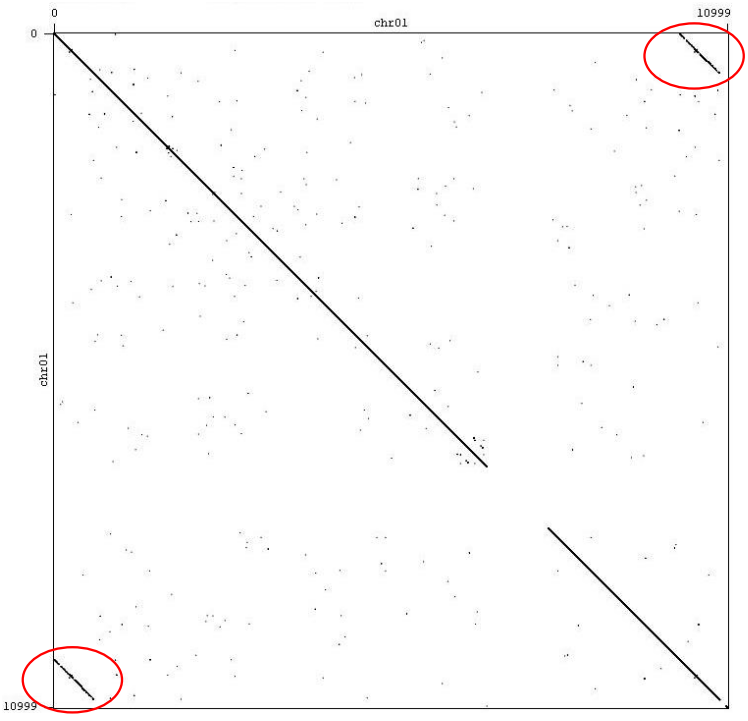


Figure 25 : Dot-plot de l'extrémité du chromosome 1 impliqué dans les deux paires de zones discordantes identifiées. Le dot-plot a été réalisé en comparant l'extrémité du chromosome 1 de la base 1 à 10 000 contre elle-même. Les cercles rouges indiquent une région dupliquée aux extrémités de la séquence comparée. Ce type de structure est caractéristique des régions LTR (Long Terminal Repeat) des retro-éléments à LTR.

II.3 Application du pipeline de détection de variations structurales pour l'analyse des accessions 'Pahang' et 'PKW'.

II.3.1 Analyse de l'accession 'Pahang' avec comme référence la nouvelle version de l'assemblage du génome A.

La recherche de variations structurales dans le génome de 'Pahang' qui pourraient expliquer les distorsions de ségrégations impliquant les chromosomes 1 et 4 a été réalisée en utilisant la nouvelle version de la séquence de référence du génome A. Pour cela j'ai exploité des données de séquencage de l'accession 'Pahang' sous la forme de deux banques :

- la banque (ensemble de lectures pairées issue du séquençage des extrémités d'un grand nombre de séquences ADN) de 5 kb présentant une couverture de 60x décrites dans le point II.2.
- et une banque de 20 kb produites sur un séquenceur illumina HiSeq2000 chez Eurofins (http://www.eurofinsgenomics.eu/) pour une couverture de 2.5x.

Le pipeline a été appliqué en utilisant tout d'abord la banque de 5 kb. Aucune signature de variation structurale n'a pu être identifiée directement par le pipeline. Toutefois, l'inspection des lectures discordantes identifiées dans les régions fortement distordues des chromosomes 1 et 4 a révélé deux types de liens :

(i) deux régions distantes de 1 Mb sur le chromosome 1 sont reliées par deux types de lectures discordantes présentant une configuration 'forward-forward' et 'reverse-reverse' respectivement (**Figure 24, flèche violette**). Ces liens pourraient être de premier abord interprétés comme une inversion de la région localisée entre ces deux zones de discordance. Cependant, une investigation fine de la sortie du pipeline montre divers indices qui suggèrent que cette détection pourrait être artéfactuelle. On observe que les paires de zones discordantes se chevauchent mutuellement alors qu'en cas de variation structurale de grande taille celles-ci doivent être contiguës mais non chevauchantes. Par ailleurs on observe une couverture plus élevé qu'attendue dans ces zones semblant indiquer que ces régions sont répétées. La comparaison de ces deux zones montre de fortes similitudes. Une étude plus fine de la structure d'une de ces zones par dot-plot en utilisant le logiciel Gepard (Krumsiek et al., 2007) a révélé une structure de type rétro-élément avec une région dupliquée au début et en fin de la séquence (**Figure 25**) typique d'un LTR (Long Terminal Repeat). Par ailleurs, la recherche de similarité avec des séquences répétées en utilisant le logiciel CENSOR (Kohany et al., 2006) a révélé que les deux zones possèdent une similarité avec des rétro-

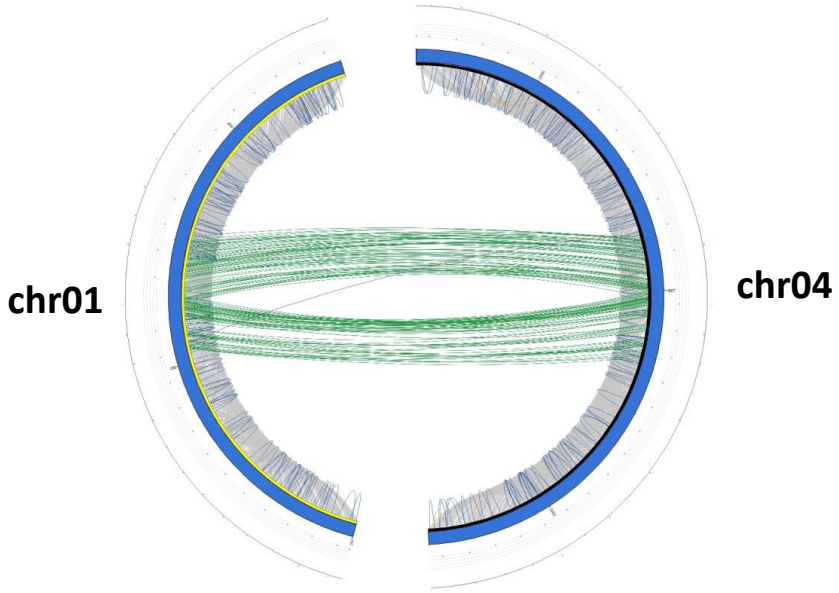


Figure 26 : Représentation Circos des lectures discordantes identifiées par alignement de la banque 5 kb de 'Pahang' sur une région de plus ou moins 10 kb autour de la paire de zone discordante identifiée entre les chromosomes 1 et 4 du génome de référence (*Musa acuminata*). Le code couleur des lectures pairées est sensiblement le même que celui des paires de zones discordantes (fr : bleu, del : rouge, ff : vert, rr : violet) avec deux différences : les lectures de type ins (bonne orientation mais « insert » trop petit) sont de couleur orange et les lectures ayant la bonne orientation et taille « d'insert » sont représentées en gris.

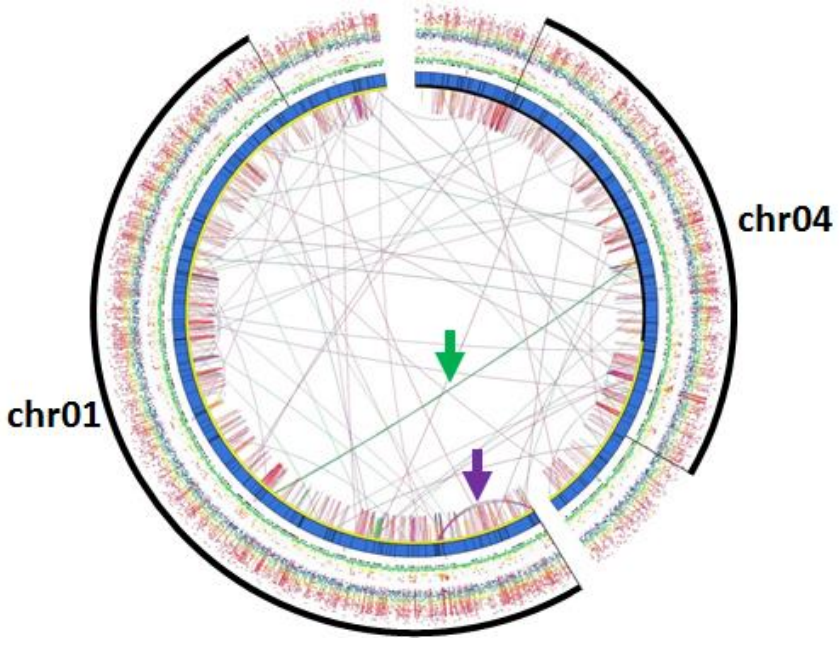


Figure 27 : Représentation Circos des lectures pairées discordantes identifiées par alignement de la banque 20 kb de 'Pahang' sur les régions les plus distordues des chromosomes 1 et 4 du génome de référence (*Musa acuminata*). Les zones en noir représentent les régions les plus distordues des chromosomes 1 et 4. La flèche violette indique deux types de lectures pairées discordantes à l'extrémité du chromosome 1. La flèche verte indique des lectures pairées discordantes reliant les chromosomes 1 et 4.

éléments à LTR de type Gypsy. Enfin, la zone liée à l'extrémité du chromosome 1 correspond exclusivement à un scaffold de petite taille (environ 10kb) dont la localisation à l'extrémité du chromosome 1 est incertaine compte tenu du peu de recombinaisons dans cette zone. Compte tenu de l'ensemble de ces éléments, il probable que les liens détectés relèvent d'un problème d'assemblage lié à la nature répétée de ces régions. L'inspection des lectures pairées issues de la banque 5kb de 'Pahang HD' alignée sur l'assemblage du génome A (l'assemblage de 'Pahang HD') révèle également des liens entre ces deux paires de zones discordantes ce qui conforte cette hypothèse de problème d'assemblage.

(ii) une région du chromosome 4, aux alentours de 27 Mb, est reliée à une région du chromosome 1, aux alentours de 3 Mb, (**Figure 24, flèche verte**) avec des lectures pairées présentant une configuration 'forward-forward'. L'inspection par dot-plot des régions des chromosomes 1 et 4 reliées ne révèle pas de structures caractéristiques de séquences répétées. De plus, l'inspection des lectures pairées s'alignant dans une fenêtre de 20 kb autour de ces régions ne révèle pas de problèmes d'assemblage puisque le reste des lectures s'alignant dans ces zones présentent de bonnes configurations d'alignement (liens gris) (**Figure 26**). Par ailleurs, la présence de lectures pairées s'alignant de façon concordantes (liens gris) dans et aux bornes des zones de discordance suggère que, si les lectures pairées observées reflètent bien une variation structurale, cette variation n'est présente que dans un des haplotypes de 'Pahang'. L'inspection des lectures pairées issues de la banque 5kb de 'Pahang HD' alignée sur l'assemblage du génome A (donc sur lui-même) ne révèle pas de de lectures reliant ces deux zones. Ces observations soulignent bien la présence d'une différence entre 'Pahang' et 'Pahang HD'. Cependant, la présence d'un seul lien reliant les chromosomes 1 et 4 n'est pas suffisante pour conclure à une variation structurale.

Ensuite nous avons analysé de la même façon les données de la banque à 20kb. La faible couverture des séquences d'extrémités de segments de 20 kb ne permet pas au pipeline d'identifier des zones de discordance avec un nombre significatif de lectures. C'est pourquoi, pour ces données, le pipeline n'a été utilisé que pour identifier les lectures discordantes et non pour identifier des zones discordantes. La **Figure 27** représente l'ensemble des lectures discordantes dans les régions distordues des chromosomes 1 et 4. La variation importante de la couverture obtenue avec la banque de 20 kb par rapport à celle obtenue avec la banque de 5 kb illustre le moins bon échantillonnage du génome par la banque de 20 kb. Trois groupes de plusieurs lectures pairées sont observés dans les trois zones significativement liées précédemment identifiées par notre pipeline avec la banque de 5kb. Les autres

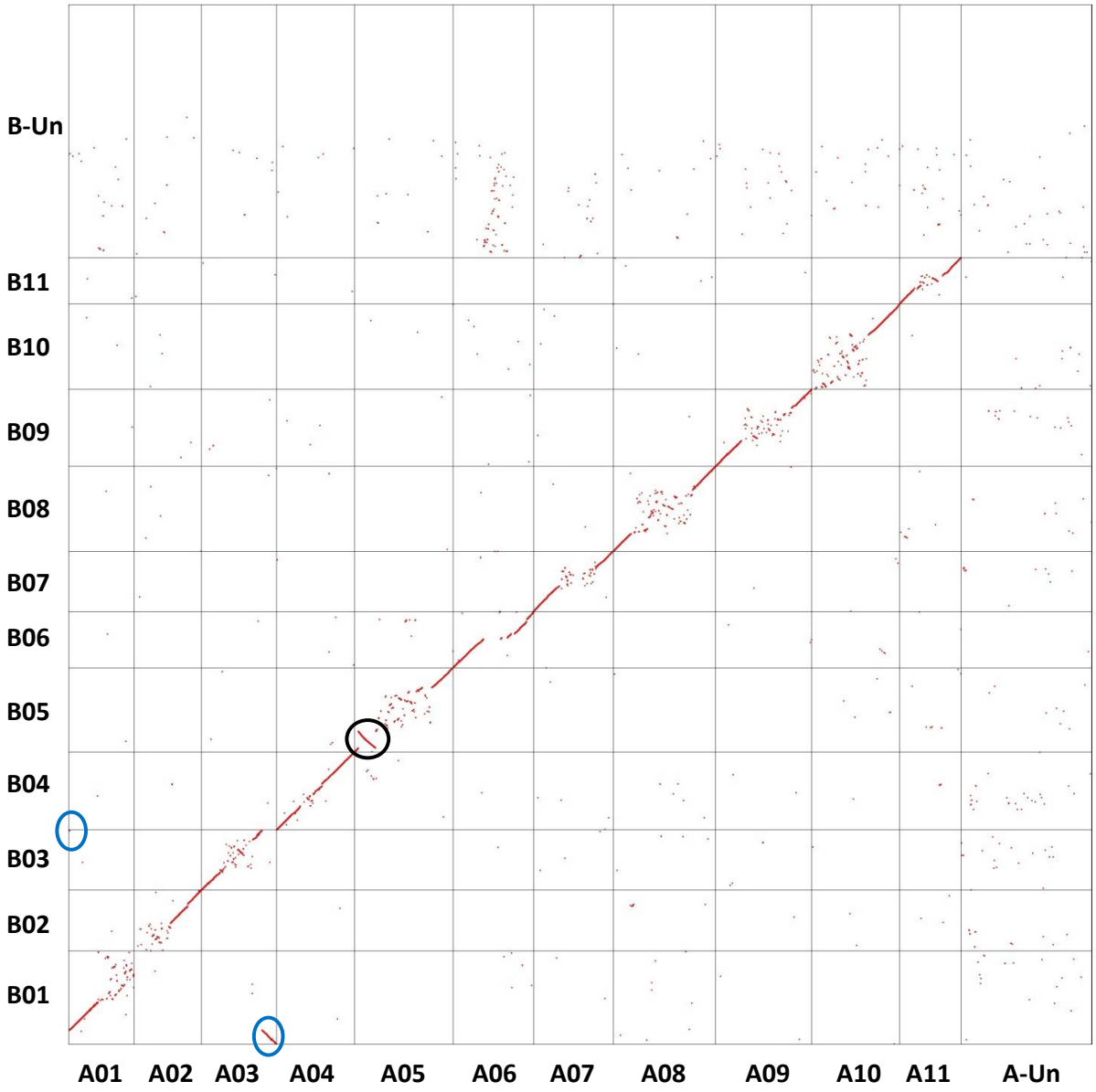


Figure 28 : Comparaison de la structure du génome de *Musa acuminata* (génome A) avec *Musa balbisiana* (génome B). Dot-plot présentant les gènes homologues entre la séquence de *Musa acuminata* (abscisses) et *Musa balbisiana* (ordonnées). Deux évènements majeurs peuvent être détectés, une inversion dans le chromosome 5 (cercle noir) et une translocation réciproque aux extrémités des chromosomes 1 et 3 (cercles bleus).

lectures pairées liant des régions du chromosome 1 avec des régions du chromosome 4 sont dispersées et ne peuvent pas être groupées pour identifier des zones de discordances. Ces lectures discordantes sont plus probablement le résultat de lectures chimériques ou de problèmes d'alignement. Ces données ne permettent donc pas d'identifier de liens supplémentaires qui permettraient de conclure quant à la présence d'une translocation.

Au final l'analyse des lectures pairées dans les zones distordues des chromosomes 1 et 4 a révélé une zone de discordance à l'état hétérozygote reliant ces deux chromosomes. Ce lien seul n'est pas suffisant pour conclure à une translocation, qu'elle soit réciproque ou non. Cependant, l'absence de détection d'une signature complète n'est pas non plus suffisante pour conclure à l'absence d'une translocation. En effet, cette absence de signature complète peut avoir plusieurs causes présentées dans la discussion générale.
Pour pouvoir conclure, le re-séquençage de l'accession 'Pahang' avec une banque à plus grand "insert" pourrait être envisagée. Par ailleurs des approches de cytogénétique moléculaire peuvent également être envisagées comme alternative ou en complément au re-séquençage.

II.3.2 Application du pipeline pour la comparaison de la structure des génomes A et B en utilisant la séquence de 'PKW' comme référence.

L'équipe est actuellement impliquée dans le séquençage *de novo* de l'accession 'PKW' appartenant à l'espèce *Musa balbisiana* (génome B) (Jin et al., in prep.). La comparaison de la séquence du génome B à la séquence de référence du génome A a mis en évidence deux différences structurales importantes (**Figure 28**). La première variation est une inversion de 9 Mb dans le chromosome 5 et la seconde est une translocation réciproque impliquant les extrémités des chromosomes 1 et 3. Pour ce dernier évènement les fragments transloqués font respectivement 8 Mb et 700 kb.
Quand cette séquence a été disponible, nous avons testé notre pipeline de recherche de variations structurales en utilisant le génome B comme séquence de référence et la banque de 5 kb générée sur l'accession 'Pahang HD'. Pour cette comparaison, le choix d'utiliser 'Pahang HD' par rapport à 'Pahang' a été motivé par le fait de choisir une accession la plus homozygote possible de façon à réduire le bruit au maximum. Notre pipeline à clairement identifié les signatures correspondant à l'inversion du chromosome 5 (**Figure 29 A**).

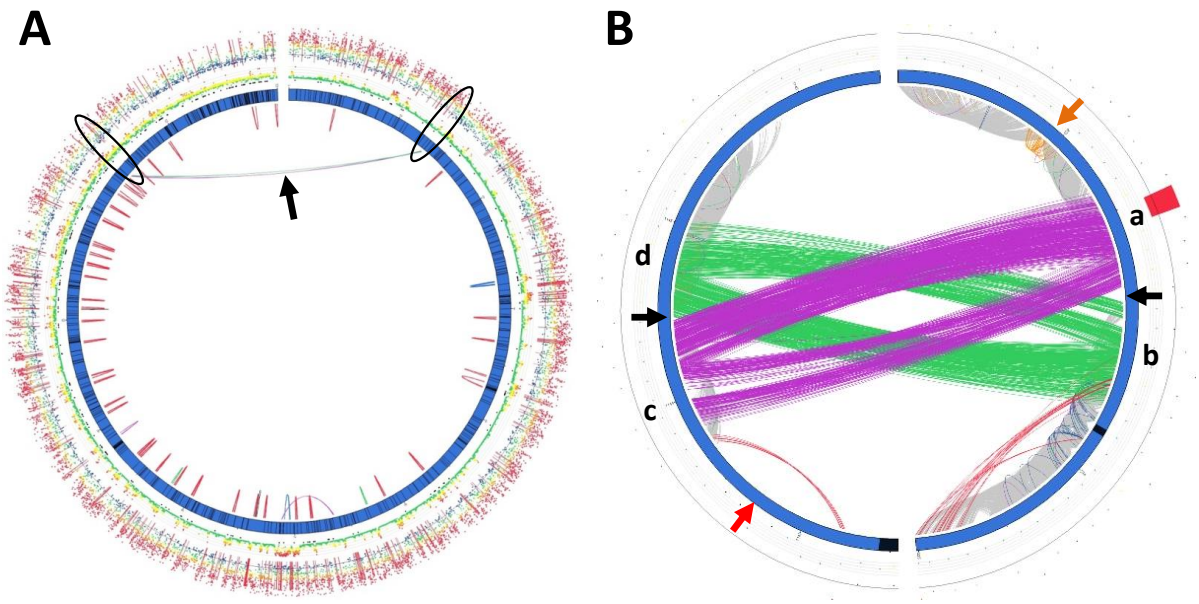


Figure 29 : Représentation Circos de la détection de l'inversion sur le chromosome 5 entre *Musa acuminata* et *Musa balbisiana*. (**A**) Représentation des paires de zones discordantes identifiées par le pipeline de recherche de variations structurales sur une région du chromosome 5 de *Musa balbisiana* comprise entre 1 et 13 Mb. Cette détection a été réalisée par alignement d'une banque 5 kb obtenue chez 'Pahang HD' et alignée sur le génome de référence *Musa balbisiana*. La configuration des paires de zones discordantes indiquées par la flèche noire permet d'identifier une inversion entre les positions 2 330 124 et 11 444 573. Les cercles noirs indiquent les points de réarrangements qui sont observés en détail en (**B**). (**B**) Représentation Circos des lectures pairées s'alignant dans une région de 10 kb autour de l'inversion détectée sur le chromosome 5. Les flèches noires indiquent les points de réarrangements. La flèche marron indique une région qui est manquante chez *Musa balbisiana* par rapport à *Musa acuminata* puisque les lectures pairées chevauchant la zone s'alignent mais avec une taille d'insert trop petite. La flèche rouge, indique une région présente chez *Musa balbisiana* mais absente chez *Musa acuminata* puisque les lectures pairées chevauchant la zone s'alignent mais avec une taille d'insert trop grande.

L'inspection des lectures pairées dans les zones de réarrangements révèle, comme attendue en cas d'inversion, deux types de lectures discordantes (**Figure 29 B**) : les lectures pairée violettes rattachent la région a et c et les lectures vertes rattachent les régions b et d. L'absence de lectures pairées reliant les régions a et b et c et d confirme que l'inversion est présente sur les deux chromosomes homologues de Pahang HD.

Par contre notre pipeline n'a pas détecté de signatures complètes correspondant à la translocation réciproque impliquant les chromosomes 1 et 3. Dans le cas d'une translocation réciproque impliquant des extrémités de chromosomes, seuls deux points de réarrangement sont attendus et donc deux types de paires de zones discordantes sont attendus (**Figure 30 A**). Seule une des paires de zones discordantes a été détectée. Nous avons cherché à connaître l'origine de la non détection de la seconde paire de zones discordantes. Pour cela nous avons étudié les lectures pairées s'alignant autour des points de réarrangements identifiés grâce aux gènes flanquant la variation structurale (**Figure 30 B**). L'inspection de ces lectures révèle un grand nombre de lectures pairées reliant les régions b du chromosome 3 et d du chromosome 1 comme attendu (**Figure 30 A**). Par contre, les lectures attendues, devant relier les régions a et c sont peu nombreuses et anormalement dispersées dans la région c (**Figure 30 B**). L'inspection des lectures pairées s'alignant dans la région c révèle peu de lectures concordantes ce qui suggère une grande divergence entre les deux génomes dans cette région. Cette forte divergence peut expliquer les problèmes d'alignement des lectures reliant les regions a et c et donc la non détection de cette variation structurale par notre pipeline. La réalisation d'un dot-plot de la région c contre elle-même révèle une structure hautement répétée. L'inspection de cette région avec le logiciel CENSOR révèle un grand nombre d'alignement sur différents types d'éléments transposables. Il semble donc que la non détection de la translocation réciproque soit due à la présence de séquence répétées au niveau d'une des bornes de la région réarrangée.

Cette comparaison des génomes A et B a été réalisée en alignant les lectures du génome A sur la séquence de référence du génome B. Il sera intéressant de réaliser l'étude inverse, c'est à dire faire cette même recherche de variations structurales en alignant des lectures pairées du génome B sur la séquence de référence du génome A. En effet, les séquences répétées aux bornes de la variation structurale chez B ont peut-être été moins bien assemblées du fait de la technologie de séquençage utilisée (Illumina) qui génère de petits fragments d'où un mauvais alignement des lectures pairées. Ces problèmes d'assemblage pourraient être moins marqués chez le génome de référence A du fait que la technologie de séquençage utilisée (454) génère des fragments plus grands.

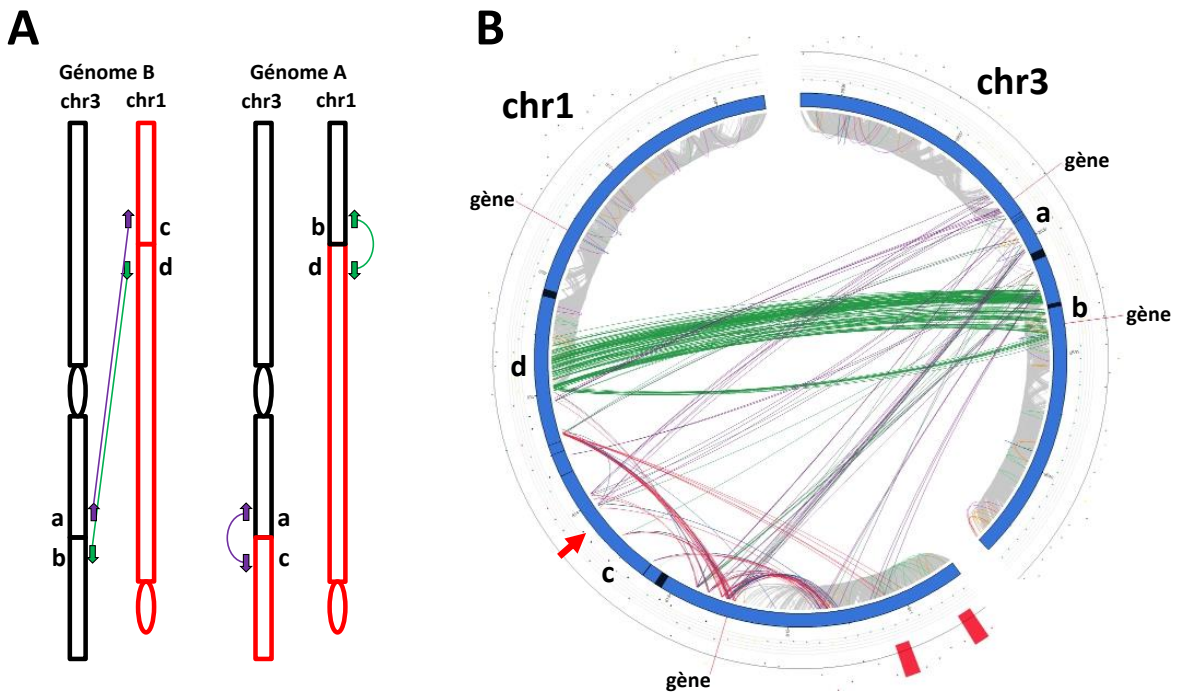


Figure 30 : Recherche de la translocation réciproque chez *Musa acuminata* et *Musa balbisiana*. (A) Schématisation de la translocation réciproque impliquant les chromosomes 1 et 3 chez *Musa acuminata* et *Musa balbisiana*. Les flèches représentent les lectures pairées obtenues sur le génome A et leur alignement attendu sur le génome B. **(B)** Représentation Circos des lectures pairées s'alignant autour des points de réarrangement de la translocation réciproque impliquant les chromosomes 1 et 3 chez *Musa acuminata* et *Musa balbisiana*. Les lectures pairées sont inspectées sur une fenêtre de 20 kb autour des points de réarrangements. Les points de réarrangements ont été identifiés par recherche de gènes homologues entre les génomes A et B. Les gènes homologues les plus proches des points de réarrangements sont localisés par le mot (gène). La flèche rouge indique une région non couverte par des lectures pairées s'alignant de façon concordante. L'absence de ces lectures ainsi que la présence de lectures discordantes indique une importante différence entre le génome A et le génome B dans cette zone.

Conclusion

Le pipeline de recherche de variations structurales développé au cours de cette thèse a été testé sur deux accessions de bananiers. Une accession proche du génome de référence (*Musa acuminata*) suspectée de présenter une hétérozygotie structurale et une accession appartenant à une espèce différente (*Musa balbisiana*) qui présente une translocation réciproque et une inversion par rapport au génome de Pahang-HD (référent actuel pour *M. acuminata*). Ce pipeline n'a pas permis de détecter la variation structurale suspectée sur l'accession 'Pahang' proche du génome de référence. Cependant il a quand même identifié une signature partielle qui pourrait correspondre à la variation structurale supposée. De plus, le pipeline a permis dans le cas de la comparaison de deux espèces dont la structure est connue, de détecter l'une des différences de structure et la signature partielle de l'autre variation.

Plusieurs raisons liées à la nature de la séquence (notamment la présence de séquences répétées) et la qualité de l'assemblage dans les régions aux bornes des réarrangements peuvent être invoquées comme limites à cette approche. Elles seront développées en discussion générale de cette thèse.

Cependant, même dans ces situations, notre pipeline peut être utile, en complément d'autres types d'informations (distorsions de ségrégation, analyses cytogénétiques moléculaires) pour identifier les variations structurales potentielles et pour mieux caractériser cette variation sur la base des signatures partielles détectées par le pipeline. Par ailleurs, il permet d'observer rapidement, au niveau des points de réarrangements, les configurations d'alignement de lectures pairées issues de re-séquençage ainsi que de visualiser différentes statistiques d'alignement (couverture et proportion de paires discordantes) et à différentes échelles.

Chapitre III : Origine des distorsions de ségrégation chromosomique chez 'Pahang' (*Musa acuminata ssp malaccencis*): variations structurales et/ou sélection ?

A ce jour, neuf cartes génétiques ont été publiées chez le bananier (D'Hont et al., 2012; Fauré et al., 1993b; Hippolyte et al., 2010; Mbanjo et al., 2012; Noyer et al., 1997; Vilarinhos, 2004). Elles concernent toutes l'espèce *Musa acuminata*. Toutes ces cartes présentent de fortes distorsions de ségrégation qui sont concentrées sur certains groupes de liaisons/chromosomes. Les parents de ces populations ségrégeantes sont issus d'hybridations entre sous-espèces de *Musa acuminata.* L'observation par cytologie des figures d'appariements chromosomiques à la méiose chez certains parents de ces populations a montré la présence de nombreux multivalents et univalents (Dessauw, 1987; Dodds, 1943; Dodds and Simmonds, 1948; Shepherd, 1999) pouvant être expliqués par des hétérozygoties structurales. L'origine de ces zones distordues a donc généralement été attribuée à leurs possibles hétérozygoties de structures chromosomiques.

Les hétérozygoties structurales sont suspectées d'être au moins en partie responsables de la faible fertilité des hybrides de bananier intersubspécifiques. D'autre part, les distorsions de ségrégation biaisent les analyses de cartographie génétique et limitent la compréhension du déterminisme génétique des caractères d'importance agronomique. Dans ce contexte, il est important de mieux comprendre l'origine des distorsions de ségrégation, de vérifier et de caractériser leur lien ou non avec différents types d'hétérozygotie structurale afin d'orienter au mieux les programmes d'amélioration variétale.

Dans le cadre du séquençage du génome du bananier (D'Hont et al., 2012), une carte génétique du parent 'Pahang' de l'haploïde doublé séquencé ('DH-Pahang') a été réalisée (cf point 3 de l'introduction). Cette carte réalisée à partir de 652 marqueurs générés sur une population de 180 individus obtenus par autofécondation de 'Pahang' (*Musa acuminata ssp malaccensis*) présente 17 % de marqueurs distordus dont la majorité a été attribuée aux chromosomes 1 et 4. L'intensité de la liaison des marqueurs distordus des chromosomes 1 et 4 est telle qu'il a été difficile d'attribuer ces marqueurs à chacun de ces deux chromosomes (D'Hont et al., 2012).

Dans le cadre de ce travail de thèse, j'ai essayé de mieux comprendre l'origine des distorsions observées dans la descendance de 'Pahang' et leur lien potentiel avec une hétérozygotie de structure. Pour cela, le génotypage de la population issue de l'autofécondation de 'Pahang' a été densifié pour réexaminer finement les profils de distorsion et mieux définir les zones distordues des chromosomes 1 et 4. D'autre part, j'ai élaboré et conduit une approche de simulation pour tester différents modèles d'interprétation qui a été appliquée sur la population 'Pahang' et utilisée pour tester les hypothèses proposées sur un croisement impliquant l'accession 'Pisang lilin' dont le profil de distorsion est similaire à celui de 'Pahang'.

Ce travail est présenté sous forme d'un projet de publication dont le titre est : **Exploring the origins of large chromosomal segregation distortions observed in a banana (*Musa acuminata*) progeny ;** et de données complémentaires.

Projet de publication n°3

Title

Exploring the origins of large chromosomal segregation distortions observed in a banana (*Musa acuminata*) progeny

Guillaume Martin[1], Françoise Carreel[1], Andrzej Kilian[2], Mathieu Rouard[3], Frédéric Bakry[1], Franc-Christophe Baurens[1], Angélique D'Hont[1*]

[1]. CIRAD (Centre de coopération Internationale en Recherche Agronomique pour le Développement), UMR AGAP, F-34398 Montpellier, France

[2]. Diversity Arrays Technology, Yarralumla, Australian Capital Territory 2600, Australia.

[3]. Bioversity International, Parc Scientifique Agropolis II, 34397 Montpellier Cedex 5, France.

[*]. Corresponding author: Angelique D'Hont

UMR AGAP, CIRAD, TA A-108/03, Avenue Agropolis, 34398 Montpellier cedex 5, France

Phone: +33 (0)4 67 61 59 27

Fax: +33 (0)4 67 61 56 05

email address: angelique.d'hont@cirad.fr

Abstract

Musa acuminata has been divided into sub-species that have diverged following geographical isolation. Hybrids between these sub-species show fertility loss, abnormal meiosis, and distorted chromosome segregation. These characteristics are attributed to chromosome structural rearrangements that occurred following geographical isolation. In the present study we analyzed the segregation of 7,417 SNP markers in a self-progeny of the wild diploid *Musa acuminata* 'Pahang' to better understand, the origin of the strong segregation distortions, pseudo-linkage and low recombination involving part of chromosome 1 and 4 previously detected in this population. Four models (duplication, reciprocal and non-reciprocal translocation and/or gametic selection) were simulated and tested. The duplication model could be excluded. The translocation models cannot be excluded but need to be associated to additional gametic selection. A model without structural heterozygosity but involving selection between 4 genetics factors may also explain the data.

Additional data will be necessary to conclude for the case of 'Pahang'. However, the methodology proposed led to a much better resolution of the distorted regions. This method allowed identification of selected haplotype combinations and values of gametic selection for each model. This methodology could help characterizing other banana or plant populations displaying strong distorted segregation.

Introduction

Reproductive isolation is both an indicator of speciation and a mechanism that maintains integrity of diverging species (Ouyang and Zhang, 2013). Several reproductive isolation mechanisms have been identified to explain the limited gene flow between diverging populations (Dobzhansky, 1937; Mayr, 1942). According to the stage when reproductive isolation arises, it can be categorized into prezygotic reproductive isolation and postzygotic isolation (Ramsey et al., 2003). Prezygotic reproductive isolation prevents the formation of hybrid zygotes through mating discrimination between divergent populations. Several mechanisms can lead to prezygotic reproductive isolation and can be classified in five groups: temporal isolation, habitat isolation, behavioral isolation, mechanical isolation and gametic isolation. Postzygotic isolation groups a series of mechanisms occurring after mating, resulting in decreased hybrid fitness. Among these mechanisms one can identify hybrid necrosis or weakness, hybrid sterility and lethality in progenies.

The *Musa acuminata* species (2n=2x=22) has been divided into sub-species that have diverged following geographical isolation in distinct South East Asia continental regions and islands (Perrier et al., 2011). Human migrations, probably during Holocene, have allowed contact between these subspecies through the transport of material (Perrier et al., 2011). This resulted in the emergence of intersubspecific hybrids with reduced fertility (Dodds and Simmonds, 1948; Fauré et al., 1993a; Shepherd, 1999). This reduced fertility in hybrids can be interpreted as postzygotic reproductive isolation between subspecies. Early farmers would then have selected hybrids producing fruits with high flesh and low seed content, two important characters for edibility. Chromosomal pairing at meiosis in *Musa acuminata* is generally regular in bivalents within accessions deriving from crosses within single subspecie, but irregular with multivalents and univalents in hybrids between sub-species. (Dessauw, 1987; Dodds, 1943; Dodds and Simmonds, 1948; Fauré et al., 1993a; Hippolyte et al., 2010; Shepherd, 1999; Vilarinhos, 2004). Structural heterozygosity has been invoked to explain those irregular pairing configurations (multivalents, univalents) in intersubspecific hybrids. This led to the classification of *Musa acuminata* accessions into seven translocation groups, named according to their geographic origins: Standard, Northern Malayan, Northern 1, Northern 2, Malayan Highland, Javanese and East African (Shepherd, 1999). Within each group, wild accessions share the same chromosomal structure and are reported as structurally homozygous while hybrids between groups are structurally heterozygous. The Standard group is the largest and includes *Musa acuminata* sub-species *microcarpa*, *banksii* and most

Table 1: Synthesis on publised maps in *Musa* genus

cross	map	*Musa acuminata* sub-species	population size	marker number	distorted markers (%)	LG number	most distorted LG	cytogenetical data	references
sef-fecondation	SFB5	hybrid[1]	92	77	36%[a]	15	SFB5-LG12	2 reciprocal translocations	Fauré et al., 1993b
sef-fecondation	M53	hybrid[2]	89	185	26%[b]	18	M53-LG9, M53-LG7	-	Noyer et al., 1997
self-fecondation	CAM	hybrid[3]	154	120	59%[a]	14	CAM-LG1, CAM-LG2, CAM-LG5	2 non reciprocal translocations	Vilarinhos, 2004
TMB_{2x} 6142-1 x TMB_{2x} 8075-7	TMB_{2x} 6142-1	hybrid[4]	81	231	14.5%[b]	15	A11, A12, A14	-	Mbanjo et al., 2012
6142-1-S x TMB_{2x} 8075-7	6142-1-S	*burmannica*	58	152	8.2%[b]	16	B2	-	Mbanjo et al., 2012
multiparental	TMB_{2x} 8075-7	hybrid[5]	139	316	40.6%[b]	15	C1, C6, C7, C9	-	Mbanjo et al., 2012
Bornéo x Pisang lilin	Bornéo	*microcarpa*	180	261	12%[b]	11	chromosomes6, 8	two structural polymorphism	Hippolyte et al., 2010
Bornéo x Pisang lilin	Pisang lilin	*malaccensis*	180	359	24%[b]	11	chromosomes 1, 2, 3, 4, 10	1 structural variation + 1 inversion	Hippolyte et al., 2010
self-fecondation	Pahang	*malaccensis*	180	652	17%[b]	11	chromosomes 1, 4	-	D'Hont et al., 2012

[a]: significance deviation from expected medelian segregation $\chi 2$ $p < 0.05$

[b]: significance deviation from expected medelian segregation $\chi 2$ $p < 0.005$

[1]: hybrid genotype containing *banksii* sub-species component

[2]: hybrid genotype containing *malaccensis* and *banksii* sub-species component

[3]: hybrid genotype containing *burmannicoides* and *banksii* sub-species component

[4]: hybrid genotype containing *burmannica* sub-species component

[5]: hybrid genotype containing *burmannicoides* sub-species component

malaccensis accessions while other groups are sub-species specific with the exception of *siamea* sub-species that is found in both Northern 1 and Northern 2 groups.

A few genetic mapping studies have been performed in *Musa acuminata*. All of them revealed important segregation distortions implicating each mainly a few linkage groups (LG) (Table 1) (D'Hont et al., 2012; Fauré et al., 1993b; Hippolyte et al., 2010; Mbanjo et al., 2012; Noyer et al., 1997; Vilarinhos, 2004). For some of these studies, chromosomal pairings at meiosis in the parents were observed and were always perturbed, suggesting the presence of chromosomal structural heterozygosities. Most of these maps were very partial, with a large number of small linkage groups. However, two of them, described below, were more saturated with 11 linkages groups, allowing deeper analysis.

A F1 population of 180 individuals obtained from a cross between *M. a. ssp microcarpa* 'Borneo' and a diploid accession 'Pisang lilin' related to *M. a. ssp malaccensis* was used to construct two parentals and a consensus map with 489 markers (SSR and DArT) grouped in 11 linkage groups (Hippolyte et al., 2010). The map established from the 'Borneo' parent comprised 12% of skewed markers from the expected Mendelian segregation ($p<0.005$) with a higher concentration in LG6 and LG8 while the map established from the 'Pisang lilin' parent exhibited 24% of skewed markers ($p<0.005$) with distinct LGs affected (LG1, LG2, LG3, LG4 and LG10). Cytogenetic studies suggested structural heterozygosity in both parents. For 'Pisang Lilin', analysis of the marker segregation distortions through tree representations and simulation approaches and chromosome pairing configurations at meiosis were interpreted as the presence of a heterozygous duplication that involved a region of LG1 duplicated in LG4 and an inversion in LG10. For the 'Borneo' parent, no clear interpretation could be deduced.

A map obtained from the self-fertilization of 'Pahang' accession (*M. a. ssp malaccensis*) was used to assemble the *Musa* reference genome sequence (D'Hont et al 2012). In this map, 17% of the 652 markers (SSR and DArT) were skewed ($p<0.005$) with higher concentration in LG1 and LG4. Interestingly, the segregation distortions observed in 'Pahang' for LG1 and LG4 were similar as the ones in 'Pisang lilin' (Hippolyte et al., 2010) and both genotypes are related to *M. a. ssp malaccensis*. For both accessions segregations distortions observed in LG1 and LG4 were associated to markers pseudolinkage. In banana, structural heterozygosity was often hypothesized to explain chromosome segregation distortions and reduced fertility in inter-sub-specific hybrids. However, Fauré et al., (1993a) analyzed meiosis of several wild and hybrids of *Musa acuminata* and found no direct relationship between structural

heterozygosity and fertility, suggesting that additional factors could be involved including genic control.

Various mechanisms have been proposed to explain segregation distortions in plants. These mechanisms can be divided into five classes: (i) Incompatible gene combinations that generate differential microspore survival (reviewed in rice - Sweigart and Willis, 2012), differential egg cell survival (reviewed in rice - Ouyang and Zhang, 2013) and differential hybrid survival (Takumi et al., 2013). (ii) Differential gamete fitness that affects probability of gamete to be fertilizing (Fishman et al., 2001). (iii) Meiotic drive that subverts meiosis so that particular chromosomal regions are preferentially found in gametes (Birchler et al., 2003; Buckler et al., 1999). (iv) Absence of a functional gene (Bikard et al., 2009). (v) Structural heterozygosity that generates irregular chromosome segregation at meiosis (Jáuregui et al., 2001). Depending on the stage when these distinct mechanisms arise will lead to gametic selection or zygotic selection.

Seedless fruits are required to produce edible banana but are a severe constraint for banana breeding. In addition, skewed segregations hamper genetic mapping analysis, limiting progress in understanding of genetic determinism of agronomic traits. A better understanding of mechanisms responsible for segregation distortions and low fertility is thus essential to help orient banana breeding strategies. At a broader level, relation between chromosome rearrangements and speciation has been a major question since Dobzhansky and Sturtevant, (1938) (Rieseberg, 2001; White, 1962).

The availability of a banana reference genome sequence (D'Hont et al., 2012) allows considering new approaches to characterize these mechanisms. In this paper, we combined highly saturated map from the *Musa acuminata* 'Pahang' progeny, genome re-sequencing approaches and simulations to investigate the origin of the segregation distortions observed in 'Pahang'.

Material

A population of 180 individuals (AF-Pahang) from the self-fertilization of the diploid accession 'Pahang' ITC609 that belongs to *Musa acuminata* subspecies *malaccensis* was produced. 'Pahang' accession is the parent of the doubled-haploid ('DH-Pahang') that was used to produce the *M. acuminata* banana reference genome sequence. This population was used to anchor the *M. acuminata* reference genome sequence assembly to the 11 *Musa* chromosomes (D'Hont et al., 2012; Martin et al., in prep.). The cross was made at the CIRAD research station in Guadeloupe, French West Indies. Embryo rescue was performed as described in (Bakry, 2008) to minimize the loss of individuals. The diploid level of the 180 individuals was checked by flow cytometry as described in Hippolyte et al., (2010). Total DNA was isolated from fresh expanding leaves of 2 month-old individuals grown in a greenhouse using a modified MATAB method (Risterucci et al., 2000). A total of 9,968 SNP markers were generated on the AF-Pahang population using the DArTseq genotyping by sequencing (GBS) technology (Cruz, 2013).

Total DNA of 'Pahang' accession was used to construct 5 kb insert size mate-pair library that was sequenced with the illumina HiSeq 2000 platform at GENOSCOPE http://www.genoscope.cns.fr. A total of 166,732,973 paired reads were obtained representing an estimated coverage of the DH-Pahang genome (523 Mb) of 61x.

Methods

--***Analysis of 'Pahang' meiotic configurations***

Meiotic configurations of 'Pahang' were analyzed according to Hippolyte et al., (2010), in meiocytes of immature anthers excised from male flower buds. A total of 27 cells and eight cells were observed at metaphases I and anaphase I, respectively, under brightfield or differential interferential contrast analysis.

-- Marker analysis

SNP marker filtering

SNP markers were filtered to reduce technical genotyping error rate (estimated around 5% in the dataset, supplemental figure S1) as described in Spindel et al, 2013 for GBS data. Briefly, i) only markers that could be located on the 11 pseudo-molecules of the Pahang reference

sequence assembly (Martin et al in prep.) were considered; ii) Assuming that recombination could not occur several times in a small window of contiguous markers (which is true for a high number of well spread markers), markers accounting for multiple recombination breakpoints, in more than 10 percent of the progeny, were discarded. This step was automatically performed with the home-made-program "GBS_corrector" available on the Southgreen platform under "scaffhunter tool" (Martin et al., in prep.).

Estimation of recombination rate

Because genotyping errors lead to artefactual recombination breakpoints, to calculate recombination rate genotyping error should be close to 0%. Genotyping data were thus corrected based on the same principle applied on marker filtering. If the genotype of an individual for a marker is different from the strict consensus of the 6 surrounding markers (3 before and 3 after), the genotype of this individual was converted to the consensus genotype. This correction step was automatically performed with program "GBS_corrector". Finally, recombination rates were estimated on a sliding windows of 500 kb along the 11 chromosomes from the corrected dataset.

Segregation distortion

All selected markers were tested for significant deviation from expected 1/2/1 Mendelian segregation ratio using a chi-square test. The $-\log_{10}$ (p-value) of the chi-square was used to represent segregation distortion levels along the 11 chromosomes.

--Graphic representations of marker linkage

Marker linkage between chromosomes

Linkage between a subset of 3,708 markers randomly selected among the 11 chromosomes was calculated using JoinMap 4.1 software (van Ooijen, 2011) with an "outbreeder full-sib family" (CP) population type. The statistics retained to describe linkage was the 'logarithm of the odd' score (LOD). Linkage between markers located along the 11 chromosomes was drawn on a warm/cold color-coded dot-plot based on the LOD statistics using "pwd2figure" script available on the Southgreen platform under "scaffhunter tool" (Martin et al., in prep.).

Marker linkage between chromosomes 1 and 4

For markers belonging to chromosomes 1 and 4, pairwise recombination frequencies were calculated using JoinMap4.1 and converted into distances using the Kosambi mapping function, as proposed in Hippolyte et al., 2010. Mapping distances were imported to DARwin 6 software (Perrier and Jacquemoud-collet, 2006) and a weighted neighbor-joining tree was built.

In addition to allow treating haplotypes (homologous chromosomes) separately, each codominant marker, belonging to chromosomes 1 and 4 was re-coded as two dominant markers. LOD scores were calculated for the co-dominant and the 'recoded dominant' markers separately.

-- Simulation analysis

Testing possible scenarios to explain chromosome 1 and chromosome 4 segregation distortions and pseudolinkage

Segregation distortions and pseudo-linkage may result from different mechanisms implying structural heterozygosities and/or gene selection. Genotypes proportions expected within and at the boundaries of various types of structural variations or selected regions can be expressed as a function of survival probability of each gamete.

Genotype proportions expected for two loci A and B located on two distinct chromosomes were put in equations. Case without structural variation versus cases with various hypothesized structural heterozygosities (duplication, translocation and reciprocal translocation) were investigated (Supplemental Figures S2, S3, S4, S5). Setting survival probability of each gamete as variables, these equations expressed (i) the expected genotype frequencies that should be observed at boundaries of a structural variation or selected regions for each of the two loci A and B: (p(AA), p(Aa), p(aa), p(BB), p(Bb) and p(bb)). (ii) Combination of genotypic frequencies expected at boundaries of structural variation or selected regions of two loci belonging to distinct chromosomes (p(AABB), p(AaBB), p(aaBB), p(AABb), p(AaBb), p(aaBb), p(AAbb), p(Aabb) and p(aabb)). For the non-reciprocal translocation hypothesis, four additional equations expressing genotype frequencies of a locus X located within translocated fragment (p(X), p(Xx), p(x) and p(.)) were added (supplemental Figure S4). For the reciprocal translocation hypothesis, eight additional equations were necessary to express genotype frequencies in locus X and Y located within translocated fragments (p(X), p(Xx), p(x), p(ØX), p(Y), p(Yy), p(y) and p(ØY)) (supplemental Figure S5). The duplication hypothesis need three additional equations to

express genotype frequencies that should be observed in a locus X located within translocated fragments (p(X), p(Xx), p(x)) (supplemental Figure S3). In this last case, the absence of tri-allelic SSR markers in the distorted region (D'Hont et al., 2012) would lead to hypothesise a recent duplication if any. In this context, the X' region is the same as the X region in supplemental Figure S3 and genotypes frequencies p(X') and p(X) could not be distinguished in the real dataset and were treated as p(X) in the equation. The same observation can be drawn for genotypes frequencies p(X'x) and p(Xx) that were treated as p(Xx) in the equation.
With the assumption of a gametic selection, variables of the equation system were an octuplet (p_1, p_2, p_3, p_4, q_1, q_2, q_3, q_4) where p_n and q_n stand for the 'survival probability' of paternal and maternal gamete respectively and 1, 2, 3, 4 stand for the four possible gamete types.
For each type of hypothesised chromosomal structure, 10,000 octuplets verifying the observed genotypic frequencies in the most distorted regions of chromosome 1 and 4 in the 'Pahang' population were randomly searched. An octuplet was kept if, for each calculated genotype proportion, the deviation from the expected is less than 0.05. Paternal gamete 'survival probability' values were searched between 0 and 1. Maternal values (q_n) were estimated from a population of 57 individuals obtained from cross between 'Pahang' and a tetraploid issued from the chromosomes doubling of the parthenocarpic 'Chicame' accession. Most probable remaining quadruplet (p_1, p_2, p_3, p_4) were then identified based on their mean value.

Simulating population

A first set of two chromosomes comprising 1,000 markers equally distributed along the two chromosomes was generated. The simulated chromosomes 1 (acrocentric) and 4 (metacentric) comprised respectively 340 and 660 SNP markers. A second set of chromosomes was derived from the first but with a particular structural rearrangement. A segregating population of 180 individuals was then generated using a Python custom script on the basis of a self-fertilization of the simulated chromosomes. An average of three recombinations per meiosis (one per chromosome arms) was allowed. A genotyping error of 5% was introduced. The gametes generated were selected following the most probable octuplets associated to each simulated dataset (chromosomal structure). For each simulated dataset, the 1,000 co-dominant markers were converted into 2,000 dominant markers to treat each haplotype separately.

Table 2: Meiotic configuratios at metaphase I in 'Pahang'

Meiotic configuration		I	II	III	IV*
N° of cell	15	-	11	-	-
	5	2	10	-	-
	3	1	9	1	
	2	-	9	-	1
	1	4	7	-	1
	1	2	8	-	1
Total No. of cell	27				

I: monovalent; II: divalent; III: trivalent; IV: tetravalent.

*: no closed tetravalent observed

Table 3: Statistics on marker density on chomosomes and recombination rate

linkage group	nb_mark	mark_nb/100kb	nb_recomb	recombination rate (nb_recomb/MB)	distorted markers (p<0.005) (%)
chr01	550	1.89	196	0.019	100
chr02	508	1.72	393	0.037	4.3
chr03	669	1.91	691	0.055	4.8
chr04	945	2.55	734	0.055	63.2
chr05	618	1.48	677	0.045	1.3
chr06	755	2.01	758	0.056	14.4
chr07	578	1.65	650	0.052	3.5
chr08	916	2.04	658	0.041	23.8
chr09	696	1.69	638	0.043	6.5
chr10	661	1.75	494	0.036	21
chr11	521	1.86	527	0.052	1.9
total	7417	1.87	6416	0.045	23.6

Simulation representation

For the two simulated datasets (co-dominant and dominant markers), a dot-plot was drawn using "pwd2figure" script. For datasets implicating non reciprocal and reciprocal translocations, two dot-plots were drawn representing markers ordered on both structures that have all chromosomal regions. For the dominant simulated dataset, an additional tree representing linkage between markers was drawn. The distortion level and the genotype proportion were calculated for each marker.

--Pahang sequence coverage analysis

'Pahang' paired-reads were mapped onto the *Musa acuminata* reference genome (Martin et al., in prep.) using Bowtie2 (Langmead and Salzberg, 2012). Multiple hit reads, read duplicates and single-end mapped reads were removed using "*scaffremodler*" tools available on Southgreen platform (Martin et al., in prep.).

Chromosome coverage was calculated using SAMtools (Li et al., 2009). Mean coverage on windows of 10 kb was plotted using R (http://cran.r-project.org).

Results

Exploring meiotic configurations of 'Pahang'

Regular pairing in 11 bivalents was observed in 15 cells out of 27. Other cells showed abnormal chromosome pairing with a few monovalents and/or multivalents (Table 2) The observation of tri- and tetravalent suggested chromosomal exchanges between two pairs of chromosomes in 'Pahang' accession. Because no closed tetravalent was observed, these multivalents may result either from reciprocal translocations, non-reciprocal translocations or the duplication of a segment of one chromosome in another chromosome. The occurrence of monovalents in preparations was compatible with these three hypotheses.

In final anaphase I, some daughter cells contained 10 and 12 chromosomes, illustrating the consequences of abnormal chromosome pairing on the distribution of chromosomes in the progenies.

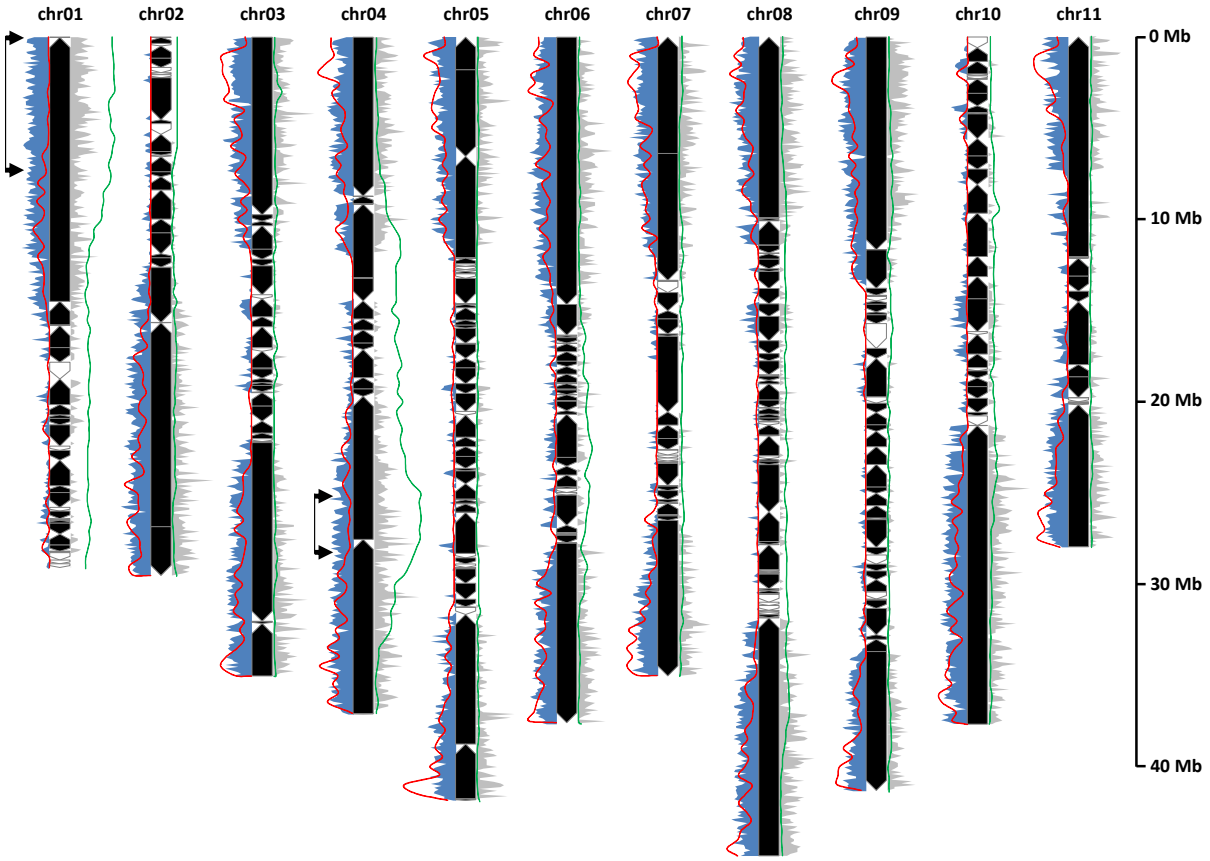


Figure 1: DArTseq marker density (grey area), gene density (blue area), segregation distortion (green curve) and recombination number (red curve) along the eleven chromosomes of *Musa acuminata*. The marker density has been calculated as the number of selected markers on window size of 100 kb along each chromosome. Gene density has been calculated as the number of bases covered by genes on windows of 100 kb. Segregation distortion has been calculated as the mean –log(p-value) of chi-square test in windows of 500 kb. Recombinations have been assessed using the corrected dataset. The recombination number has been calculated as the number of recombinations observed in the population of 180 individuals within a sliding window of 500 kb. Black arrows indicate two regions of chromosomes 1 and 4 where recombination rate is inferior to mean recombination rate and markers are strongly distorted.

Table 4: Genotypes proportions in most distorted zones

	chr01	chr04	chr06	chr08	chr10
homozygous1	0.52	0.54	0.21	0.34	0.25
heterozygous	0.40	0.38	0.64	0.53	0.60
homozygous2	0.08	0.08	0.16	0.14	0.15

chr: chromosome

Large segregation distortions in the self-progeny of 'Pahang'

A total of 9,968 SNPs were obtained on the 180 self-progeny of 'Pahang' from which, after the filtering steps, 7,417 were kept for further analyses. The markers were well distributed on the 11 chromosomes of the *Musa accuminata* reference genome assembly (Martin et al in prep.), with an average of 1.7 markers per 100 kb. As expected, a much higher marker density was observed in gene-rich regions compared to repeat-rich pericentromeric regions (Figure1, Table 3).

Recombination rate

The average recombination rate was 0.045 recombinations per Mb (Table 3). The recombination rate was in general positively correlated with gene density to the exception of a few gene-rich regions showing low recombination rate (Figure 1): a large distal region of 7.5 Mb in acrocentric chromosome 1 (from 0 Mb to 7.5 Mb) and a region of 4 Mb in the median part of metacentric chromosome 4 (from 24.5 Mb to 28.5 Mb) (Figure 1).

Segregation distortions

In total, 24% of the markers deviated from the expected Mendelian ratio (AA=0.25; Aa=0.5; aa=0.25) ($\chi 2$ test, significance $p < 0.005$) (Table 3). These markers were mainly located on chromosomes 1 and 4 which exhibited each a large region with very strong segregation distortions. Chromosomes 1 and 4 contained 100% and 63% of distorted markers respectively. Along chromosome 1, the highest distortions were observed in a region of 7.5 Mb between positions 0 to around 7.5 Mb which corresponds to the region with low recombination. In chromosome 4, the highest distortion was observed in a region of 3.5 Mb between positions 24.5 Mb to around 28 Mb, which also correspond the region with low recombination. In the most distorted regions of chromosomes 1 and 4, the segregation was biased with an excess of one homozygous genotype at the expense of the alternative homozygous and heterozygous genotypes (AA=0.53; Aa=0.39; aa=0.08 on average).

Three additional chromosomes exhibited hot spot of distorted markers but at lower extent and intensity (Figure 2). In the distorted region of chromosomes 6 and 10, marker segregation displayed an excess of heterozygous genotypes and a lack of at least one homozygous genotype (Table 4). In the distorted region of chromosome 8, marker segregation was biased with an excess of one homozygous genotype at the expense of the alternative homozygous genotype.

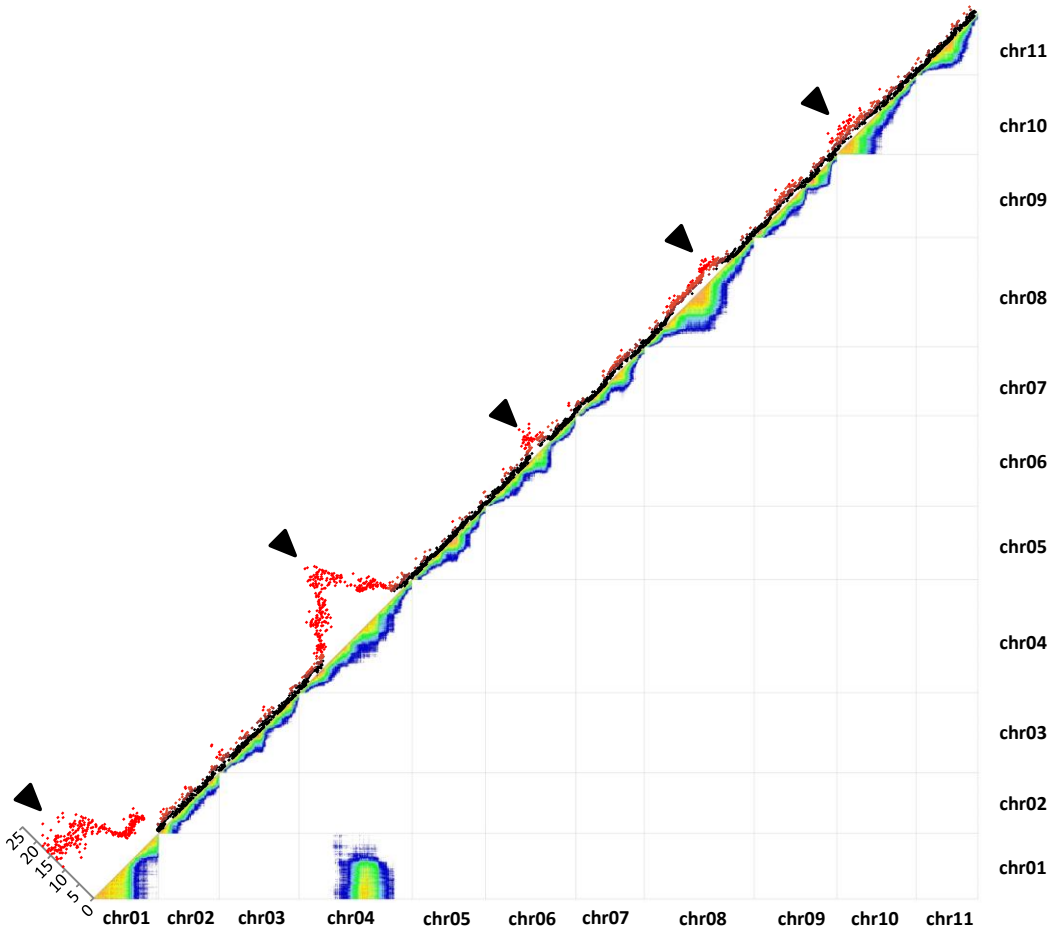


Figure 2: Dot-plot showing pairwise linkages of co-dominant markers along the eleven chromosomes. The statistics used to estimate linkage is the LOD score. Markers pairs having LOD value superior to 61, 53, 45, 38, 30, 23, 15, 8 are drawn with red, pink, orange, yellow, light green, green, light blue and blue dots respectively. When LOD value is inferior to 8 no dot is drawn. The graph represents the marker segregation distortions along the eleven chromosomes. The value presented in the graph is the $-\log_{10}$(p-value of the chi-square test testing the deviation from expected Mendelian segregation ratio).

Table 5: Chromosome 1 and 4 genotype combination frequencies (%) in the most distorted zone

Genotype combinations	Expected (medelian segregation)	Observed
1^{AA} - 4^{BB}	6.25	51.67^{+}
1^{AA} - 4^{Bb}	12.50	0.00
1^{AA} - 4^{bb}	6.25	0.00
1^{Aa} - 4^{BB}	12.50	2.22^{-}
1^{Aa} - 4^{Bb}	25.00	37.78^{+}
1^{Aa} - 4^{bb}	12.50	0.00
1^{aa} - 4^{BB}	6.25	0.00
1^{aa} - 4^{Bb}	12.50	0.00
1^{aa} - 4^{bb}	6.25	8.33^{+}

$^{+}$ over represented genotype combinations

$^{-}$ under represented genotype combinations

Pseudo-linkage

Several clustered markers of chromosomes 1 and 4 appeared strongly linked together, with a linkage intensity in the same order of magnitude than those of adjacent markers of their respective chromosomes (Figure 2 and 3). If all these markers were grouped using a classical genetic map approach, they would be clustered in a single linkage group. The attribution of markers using their location on the sequence of DH-Pahang (Martin et al., in prep.) allowed splitting this group. These markers linked the distal part of acrocentric chromosome 1 with a pericentromeric region of chromosome 4. The two linked regions were those showing strong segregation distortions and low recombination rate (Figures 1 and 2). The tree representation of these linkages led to a structure with three arms. Two arms contained markers from each arm of chromosome 4 and the third one, markers from the unique arm of chromosome 1 (Figure 2B). Linkages between haplotypes were determined by inspecting co-dominant markers re-coded into two dominant markers. The over-represented haplotypes of chromosomes 1 and 4 were found linked together (Figure 4) in the segregating population. The under-represented haplotypes were also linked together but at lower extent. This weaker linkage is due to the very low frequency of these haplotypes in the progeny (Figure 4).

Genotype combinations of chromosomes 1 and 4

Genotype combinations between chromosomes 1 and 4 in the targeted regions were inspected (Table 5). Over the nine possible genotype combinations, only four genotype combinations were found and only three genotype combinations exhibited a frequency higher than 5%. These three genotype combinations exhibited a frequency higher than the expected in case of Mendelian segregation while the last genotype combinations were under the expected (Table 5). The three over-represented genotype combinations were two genotypes that were homozygous for both chromosomes (51.67% and 8.33% against 6.25% expected) and a genotype that was heterozygous at both chromosomes (37.78% against 25% expected). The fourth present genotype combination was a combination of a heterozygous genotype for chromosome 1 and a homozygous genotype for chromosome 4. This last combination exhibited a frequency of 2.22% (against 12.5% expected). This last combination implied that marker segregations in these two regions are not completely identical and explained the slight difference observed in segregation distortions between chromosomes 1 en 4 in Table 4.

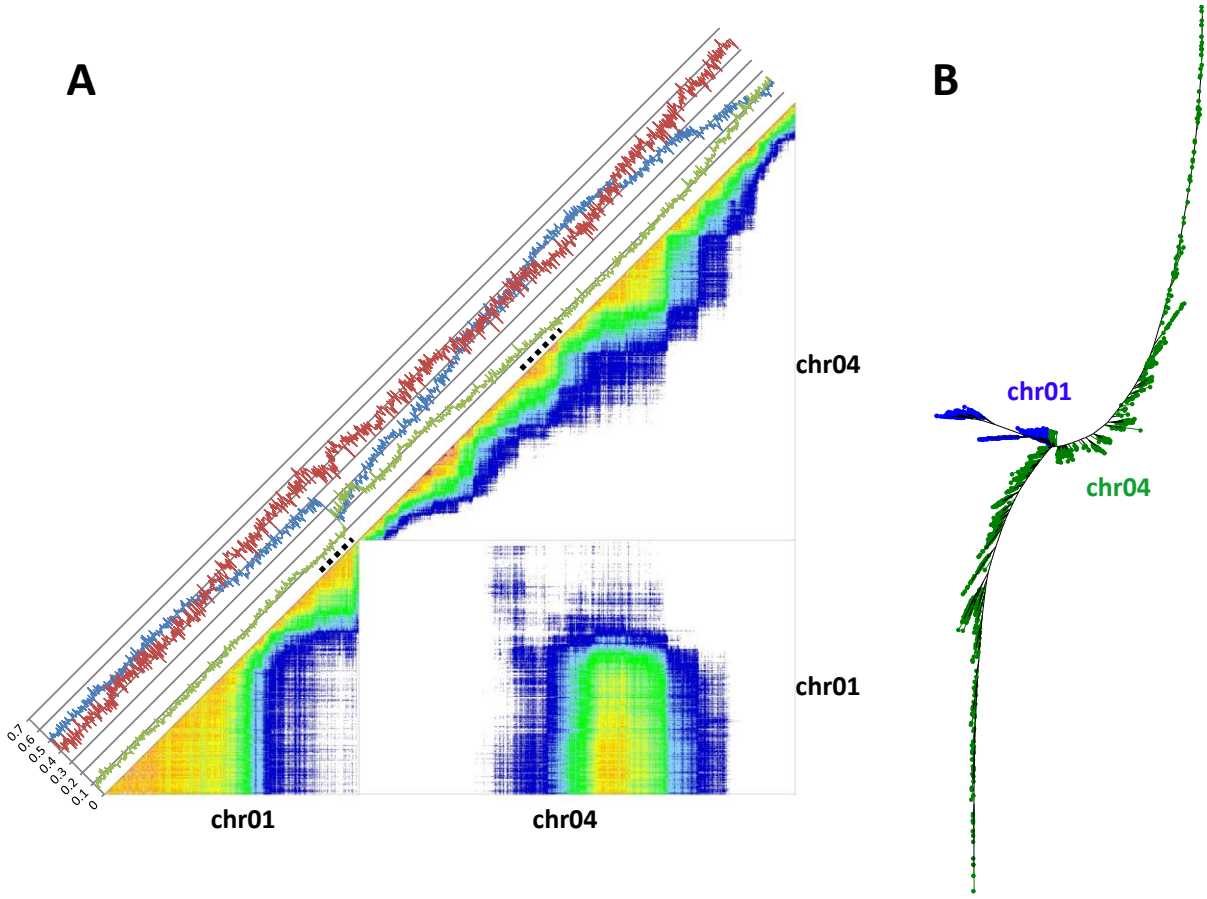


Figure 3: A) Dot-plot showing pairwise linkages of co-dominant markers along the chromosomes 1 and 4. The statistics used to estimate linkage is the LOD score. Markers pairs having LOD value superior to 61, 53, 45, 38, 30, 23, 15, 8 are drawn with red, pink, orange, yellow, light green, green, light blue and blue dots respectively. When LOD value is inferior to 8 no dot is drawn. The graph on the upper diagonal of the dot-plot represents for each marker the frequency of heterozygous, homozygous1 and homozygous2 individuals in red, blue and green, respectively. Centromeric regions are indicated with dotted lines. **B) Weighted Neighbor-joining tree constructed on the pairwise map distances using the Kosambi mapping function of marker belonging to linkage chromosome 1 and chromosome 4.** The distances were calculated from the pairwise recombination frequencies calculated by JoinMap4.1. Marker belonging to chromosome 1 and chromosome 4 are drawn in blue and green respectively.

Exploring the cause of chromosomes 1 and 4 peculiar segregation through simulation

Applying 'Pisang lilin' duplication model to the 'Pahang' population

The similar segregation distortions and pseudo-linkage observed in 'Pahang' population compared to the previously published 'Pisang lilin' genetic map (Hippolyte et al., 2010), leaded us to test the model proposed by the authors to explain our observations. This model implied a duplication of the distal part of chromosome 1 into chromosome 4 and lethality of gamete having the duplicated fragment in both chromosomes 1 and 4 at heterozygous state.

The simulation of this model revealed that genotypes frequencies were similar to those observed in 'Pahang' population (supplemental Figure S6). However, marker linkages between chromosomes 1 and 4 in the simulated population were significantly weaker than the markers linkage observed in the 'Pahang' population (Figure 3 compared to supplemental Figure S6 and Figure S7). In addition, haplotype linkage observed with simulated dominant markers revealed that over-represented haplotype of chromosomes 1 were linked with under-represented haplotype of chromosome 4 and that under-represented haplotype of chromosome 1 was linked to over-represented haplotype of chromosome 4 (supplemental Figure S7). These genotypes combinations of haplotypes between chromosome 1 and 4 did not correspond to the observed combinations in 'Pahang' population.

The duplication model with the gamete selection proposed by Hippolyte et al., (2010) thus did not explain our observations in 'Pahang' population. However, our simulation approach allowed to test other gametic selection variables that may fit our dataset in the presence of a duplication. In addition, since large segregation distortions in banana progenies have been previously attributed to translocation events (Fauré et al., 1993b; Vilarinhos, 2004), gamete survival parameters combined with other type of large structural variations were investigated.

Exploring new gamete survival variables associated with structural models to fit the observed genotype proportions

Four models were tested to explain distorted segregation and pseudo-linkage observed in chromosomes 1 and 4. The models tested were: i) gene selection without structural heterozygosity, ii) duplication of a region of chromosome 1 in a region of chromosome 4, iii) non reciprocal translocation and iv) reciprocal translocation (Figure 5, Supplemental Figures S2, S3, S4, S5). The variables of these models were the survival probability of each gamete (gametic selection) noted p_1, p_2, p_3, p_4, q_1, q_2, q_3 and q_4 where p_n and q_n stand for the 'survival

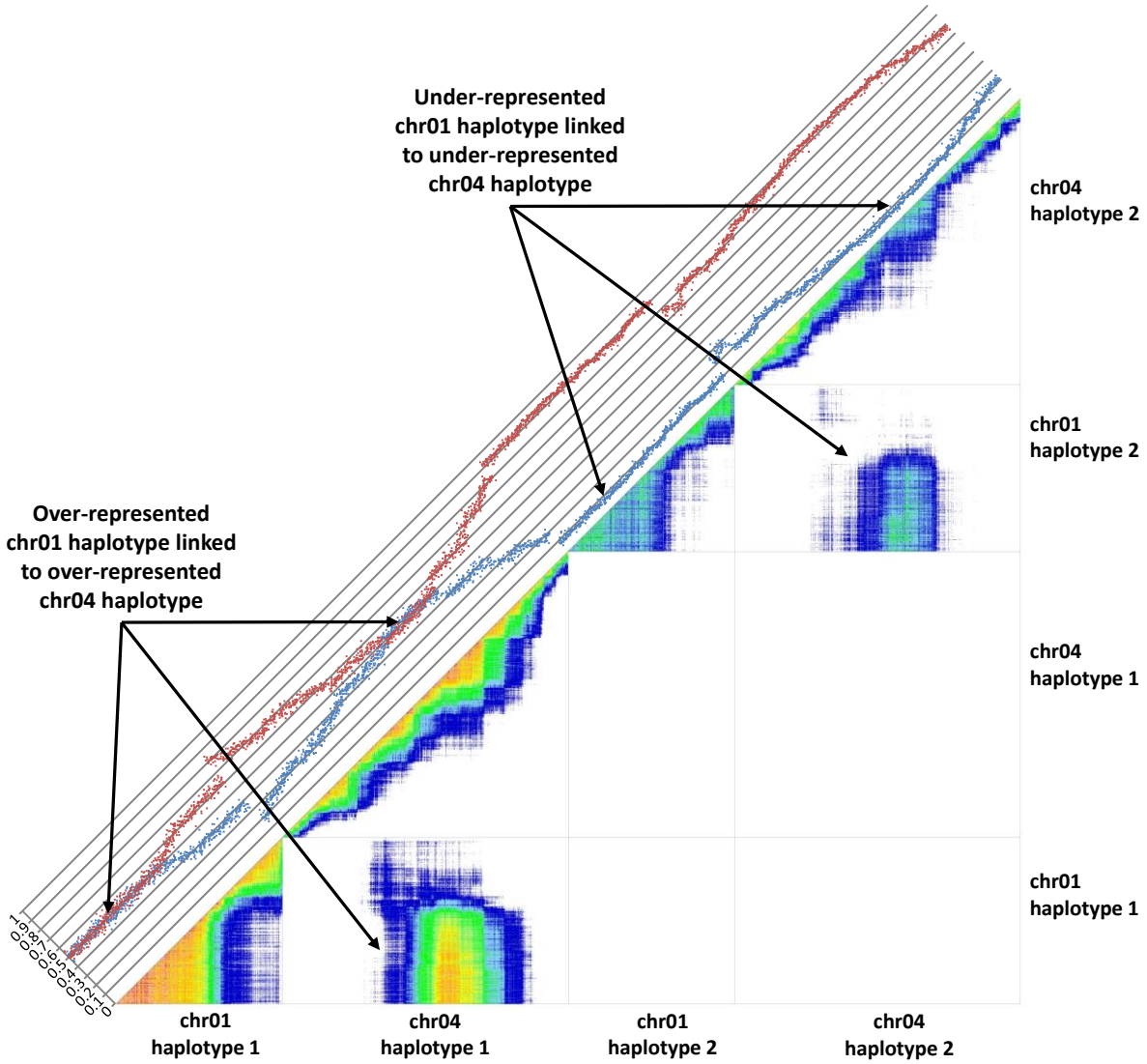


Figure 4: Dot-plot showing pairwise linkages of dominant markers along the chromosomes 1 and 4. The statistics used to estimate linkage is the LOD score. Markers pairs having LOD value superior to 61, 53, 45, 38, 30, 23, 15, 8 are drawn with red, pink, orange, yellow, light green, green, light blue and blue dots respectively. When LOD value is inferior to 8 no dot is drawn. The graph on the upper diagonal of the dot-plot represents for each marker the frequency of recessive and dominant individuals in blue and red respectively.

probability' of paternal and maternal gamete respectively and 1, 2, 3, 4 stand for the four possible gamete types.

Maternal variables of gamete survival were obtained from another progeny implicating 'Pahang' accession as maternal donor. This population was genotyped with 37 SSR markers located on chromosomes 1 and 4. The maternal survival probabilities were estimated in the most distorted regions of chromosomes 1 and 4 and values were 1, 0, 0 and 0.28 for q_1, q_2, q_3, q_4 respectively. These values were fixed for subsequent simulations and the p_n variables were searched by iterations (see material and method section). Searched equation parameters were the observed genotypes frequencies in the distorted region of chromosome 1 and 4 (Table 4) and the observed genotypes combinations in the 'most linked regions' of the chromosomes 1 and 4 (Table 5). The variation in genotype frequencies along chromosomes 1 and 4 was continuous without important shifts between contiguous markers (Figure 3). Consequently, if structural variation exists, the genotype frequencies within and at the boundaries of the duplicated/translocated fragments should be very close or equal. In this context, the genotypes frequencies observed in the most distorted region of chromosome 1 and 4 (Table 4) were used to estimate survival probability of gamete either within or at the boundaries of the duplicated/translocated fragments.

For all models, possible p_n values were relatively close (Supplemental Figures S8, S9, S10, S11) allowing calculation of a mean value for models variables. For gametic selection alone model, duplication model and simple translocation models p_1, p_2, p_3 and p_4 values were similar and respectively of 1, 0.07, 0.09 and 0.4. For the reciprocal translocation model p_3 values was slightly different with a value of 0.1. Paternal gametes survival probabilities were close to the maternal gamete survival deduced from the other progeny implicating 'Pahang' as maternal donor.

Simulating the population

The p_n, q_n values found were used to simulate the four segregating populations corresponding to the four chromosome structures tested. For the four tested models, the simulated segregation distortions and genotype proportions were consistent with observed values (Figure 3 compared to supplemental Figures S12, S15, S18 and S21).

In addition, for the four tested models, all tree drawn showed a three armed structure as observed on the real dataset (Figure 3 compared to Supplemental Figures S14, S17, S20 and S23). One exception could be observed with the tree generated on data simulating gametic selection alone (Figure S14): a small group of markers belonging to chromosome 4 generated

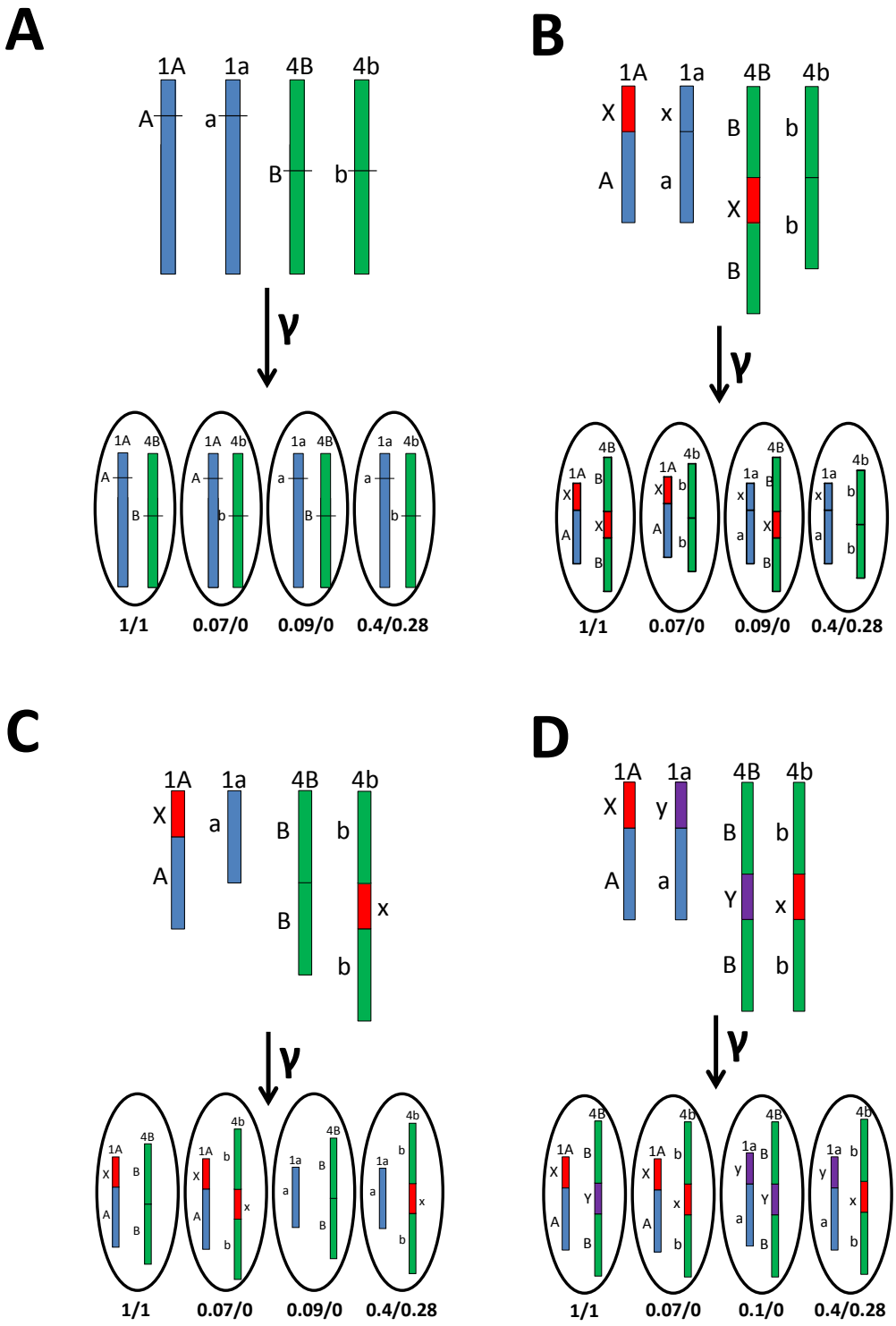


Figure 5: Schematic representation of the four tested models with zygote structure and generated gamete types. (A) Two locus selection model with selection applied on loci (A/a) and (B/b) combination. (**B**) Duplication model with a duplication of X region of chromosome 1 (in red) into chromosome 4B. (**C**) Non-reciprocal translocation model with the X region (in red) translocated from chromosome 1a to chromosome 4b. (**D**) Reciprocal translocation model with exchange of the X region (in red) and Y region (in purple) between homologous chromosomes 1 and 4. Gamete predicted survival probability is indicated under each gamete: the first value indicates paternal predicted survival and the second indicates maternal survival deduced from Pahang x ChicameT cross.

a fourth arm. All these markers correspond to one extremity of chromosome 4 and were neither distorted nor selected. This dubious grouping may highlight one limit of Neighbor joining statistics.

For all tested chromosome structures, over-represented chromosome haplotypes were linked together and under-represented chromosome haplotypes were linked together as observed in the Pahang population (Figure 4 compare to Supplemental Figure S13, S16, S19 and S22).

Comparison of both selection alone model and duplication model to real data

The dot-plots representing marker linkage under both models showed a similar pattern of marker linkage to the one observed with the real dataset (Figures 3 and 4), although slightly weaker, both for co-dominant markers and dominant markers Selection alone model dot-plots are presented in supplemental Figures S12 and S13, and duplication model are presented in supplemental Figures S15 and S16.

Comparison of non-reciprocal translocation model to real data

For the model with non-reciprocal translocation, two dot-plots representing markers linkages were drawn representing both haplotypes structure that contain all chromosome segments but in one copy (*i.e.* if the translocated segment was in chromosome 1, it is absent from chromosome 4). For one of the structure (corresponding to the over-represented haplotype), a breakpoint in marker linkages could be observed between the translocated segment and its contiguous segment in chromosome 1 (Supplemental Figures S18A and S19A). This breakpoint was not observed in the real dataset. For the other structure (corresponding to the under-represented haplotype) and under the same model, no breakpoint was observed between markers from translocated segment and their contiguous segments on chromosome 4 (Supplemental Figures S18B and S19B). Under this model, only a small region of chromosome 1 was linked to the translocated segments and this linkage was weaker than the one observed on real dataset.

Comparison of reciprocal translocation model to real data

For the model with reciprocal translocation, two dot-plots representing marker linkage were drawn representing both haplotype structures that contain all chromosome segments (*i.e.* if one translocated segment was in chromosome 1, the other translocated segment was attributed to chromosome 4). For one of the structure (corresponding to the over-represented haplotype), linkage breakpoints between markers locating in translocated segments and markers locating

in the vicinity of these translocated fragments can be observed (Supplemental Figures S21A and S22A). These breakpoints were not observed in the real dataset. For the other structure (corresponding to the under-represented haplotype), no breakpoint in marker linkage structure was observed but the linkage between translocated segments was slightly weaker than observed on real dataset (Supplemental Figures S21B and S22B).

'Pahang' re-sequencing coverage

A total of 166,732,973 'Pahang' 5 kb paired-reads were mapped onto the *Musa acuminata* reference genome. In case of duplication, an increase of coverage of 50% relative to the median coverage was expected in the most distorted part of either chromosome 1 or chromosome 4 or both. The absence of such increase (Supplemental Figures S24 A and B) suggested that these regions are not duplicated in 'Pahang' accession.

Discussion

In this paper, we tested four hypotheses to explain the distorted segregation and pseudo-linkage observed in the progeny of the *M. acuminata malaccensis* accession 'Pahang'. The four models elaborated were: i) a "gene selection alone" model, ii) a duplication model involving part of chromosome 1 into one haplotype of chromosome 4, iii) a non-reciprocal translocation model implicating a genomic region present in chromosome 1 and chromosome 4 and iv) a reciprocal translocation model with two regions exchanged between chromosome 1 and chromosome 4. For the four models tested, variables were gamete survival (i.e. probability of the gamete to contribute to a zygote). The choice of gamete survival rather than zygote survival was motivated by two observations: i) the markers segregation in the self-progeny of 'Pahang' was similar to one observed in a biparental cross between 'Pahang' (female) and the tetraploid accession 'Chicame T' (male); ii) in intersubspecific *Musa hybrids*, pollen and ovule viability is generally low (Shepherd, 1999).

We observed a slight difference between the predicted paternal and measured maternal gamete survival variables (1, 0.07, 0.09, 0.4 / 1, 0, 0, 0.28). For the first time in banana, such parameters were obtained for paternal and maternal gamete separately. The slight difference observed may be due to the relatively small size of the 'Pahang' self-progeny (180) and the

Pahang x ChicameT progeny (57) studied that do not allow observing rare chromosome 1 and 4 gametes or genotype combinations respectively. Indeed, in the Pahang x ChicameT progeny, we observed only two gamete types over the four possible types. In the 'Pahang' self-progeny, we observed only one of the six rare genotype combinations (heterozygous genotype for chromosome 1 and homozygous genotype for chromosome 4).

For the four tested model, 'unique' gamete survival variables were found and were similar for all models, consisting in nearly complete lethality of two gametes and one of the remaining gamete being much more represented than the other one (Figure 5).

The duplication model associated with only one type of gamete eliminated that was previously proposed to explain the strong distortion affecting chromosomes 1 and 4 in the progeny of 'Pisang Lilin' accession (Hippolyte et al., 2010) did not fit with our observations. In particular, it did not fit observed haplotype linkages and linkages intensity between haplotype was significantly weaker than expected. Although we found gamete survival variables that can be associated to the duplication model, the 'Pahang' re-sequenced data did not reveal significant increase of read coverage in the chromosome 1 and 4 segments involved in the distortion. The duplication hypothesis was thus excluded.

For the reciprocal and non-reciprocal translocations, observation of breakpoints in marker linkages, when markers were ordered according to over-represented haplotypes, were not congruent to our dataset. However, when markers were ordered according to under-represented haplotypes, no breakpoint in marker linkages was observed, and this models were thus congruent to markers linkage in 'Pahang' population (Figures 3 and 4 compared to supplemental Figures S18B, S19B, S21B and S22B). In this context, translocation models cannot be excluded but they implied that the 'DH-Pahang' accession that has been used to order markers correspond to the under-represented haplotype of 'Pahang' (gamete with p_4/q_4 of 0.4/0.28, Figure 5C and 5D). For the reciprocal translocation model, the low recombination observed in chromosomes 1 and 4 can be explained by reduced pairing of translocated fragments. In the model of non-reciprocal translocation, the low recombination in part of the chromosome 4 can be explained for the same reason but the low recombination in chromosome 1 arm remains to be explained.
The two loci selection model (Figure 5A) could also fit the observed data but also requires additional explanations to resolve the low recombination in chromosome 1.

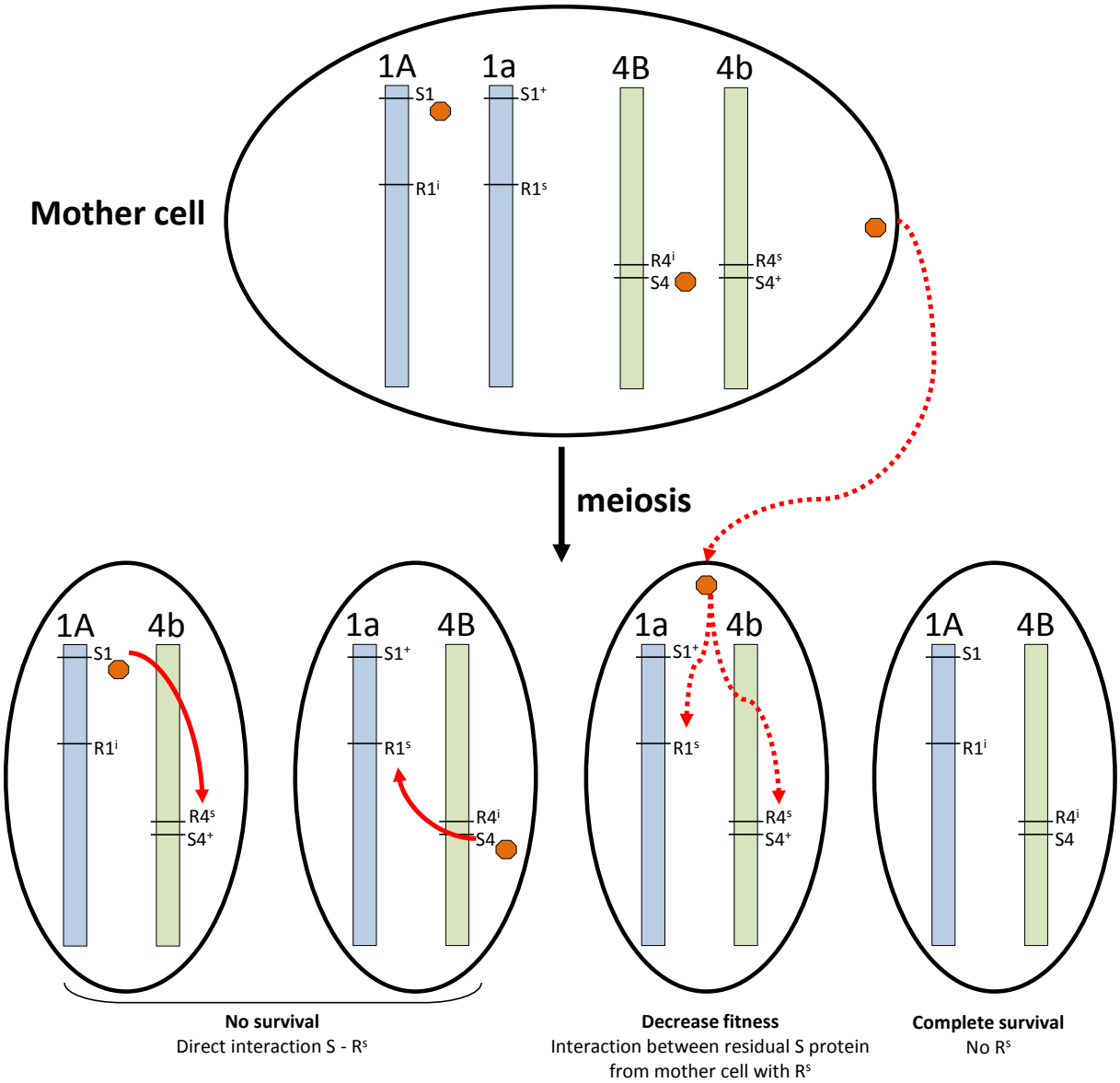


Figure 7: Hypothetical multiple loci interaction model to explain gamete survival obtained from the selection alone model. The mother cell is heterozygous S1 $R1^{i}/S1^{+}$ $R1^{s}$ in chromosome 1 and S4 $R4^{i}/S4^{+}$ $R4^{s}$ in chromosome 4. In gamete, the interaction of S alleles product with R^{s} allele is deleterious resulting in death of the two gamete types. The third gamete has its fitness reduced because of hyper-sensitivity and interaction with residual S proteins produced in mother cell. The last gamete has $R1^{i}$ and $R4^{i}$ allele and is thus not sensitive to S protein. This last gamete showed complete survival.

For the three remaining models, gamete mortality needs to be explained. With the reciprocal translocation model, the two gametes showing strong mortality are those that are missing genomic fragment (Figure 5D). The absence of such large genomic fragment implied several missing genes that are most probably deleterious for gamete.

With the non-reciprocal translocation model, the two gametes showing the strongest mortality correspond to the gamete lacking the translocated fragment on both chromosomes and the gamete having the translocated fragment on both chromosomes (Figure 5C). As mentioned before, the absence of a complete chromosome segment containing genes can be deleterious. The doubling of a chromosome segment as source of lethality is less obvious. Interestingly, the gamete having the translocated fragment on both chromosomes is also the gamete that is supposed lethal in the 'Pisang lilin' model (Hippolyte et al., 2010). One explanation for which such combination could be deleterious may be found in the gene balance hypothesis (Birchler and Veitia, 2007). The duplication of genes located in the translocated region may disrupt gene stoichiometric balance resulting in non-viability of gametes having duplicated fragments.

For both translocation models, one of the two remaining gametes showed a lower transmission ratio. The source of this lower transmission ratio could be the presence of less fit allele(s) located in or at the vicinity of the duplicated fragments. The presence of such alleles in the center of chromosome 1a in case of a non-reciprocal translocation (Figure 5C) may explain the low recombination rate observed in chromosome 1 arm.

Regarding the case of selection alone model (Figure 5A), to our knowledge no previous explanation model fit our data although examples for non-homologous chromosomes co-segregation have been reported (Bikard et al., 2009; Longley, 1945; Mizuta et al., 2010). However, a model with four loci consisting in an extension of the "Segregation Distorter" (SD) meiotic drive gene complex observed in *Drosophila melanogaster* (reviewed in Larracuente and Presgraves, 2012) may fit our model. In *Drosophila melanogaster,* the SD complex distorts spermatogenesis of SD/SD+ males that almost exclusively produce SD spermatids rather than the expected 1:1 Mendelian ratio. This complex disrupts chromatin condensation needed for spermatid formation. Mechanisms by which this complex works are not completely resolved but two key loci are involved: the driver of segregation distortion (Segregation distorter – Sd) and the target of the drive (Responder – Rsp). The product of Sd gene impacts the functionality of Rsp locus only when Rsp^{s} allele is present. In drosophila, heterozygous males that are sd Rsp^{i} /sd^{+} Rsp^{s} produced almost only sd Rsp^{i} spermatid. In banana, the duplication of such system in chromosomes 1 and 4 leads to four possible

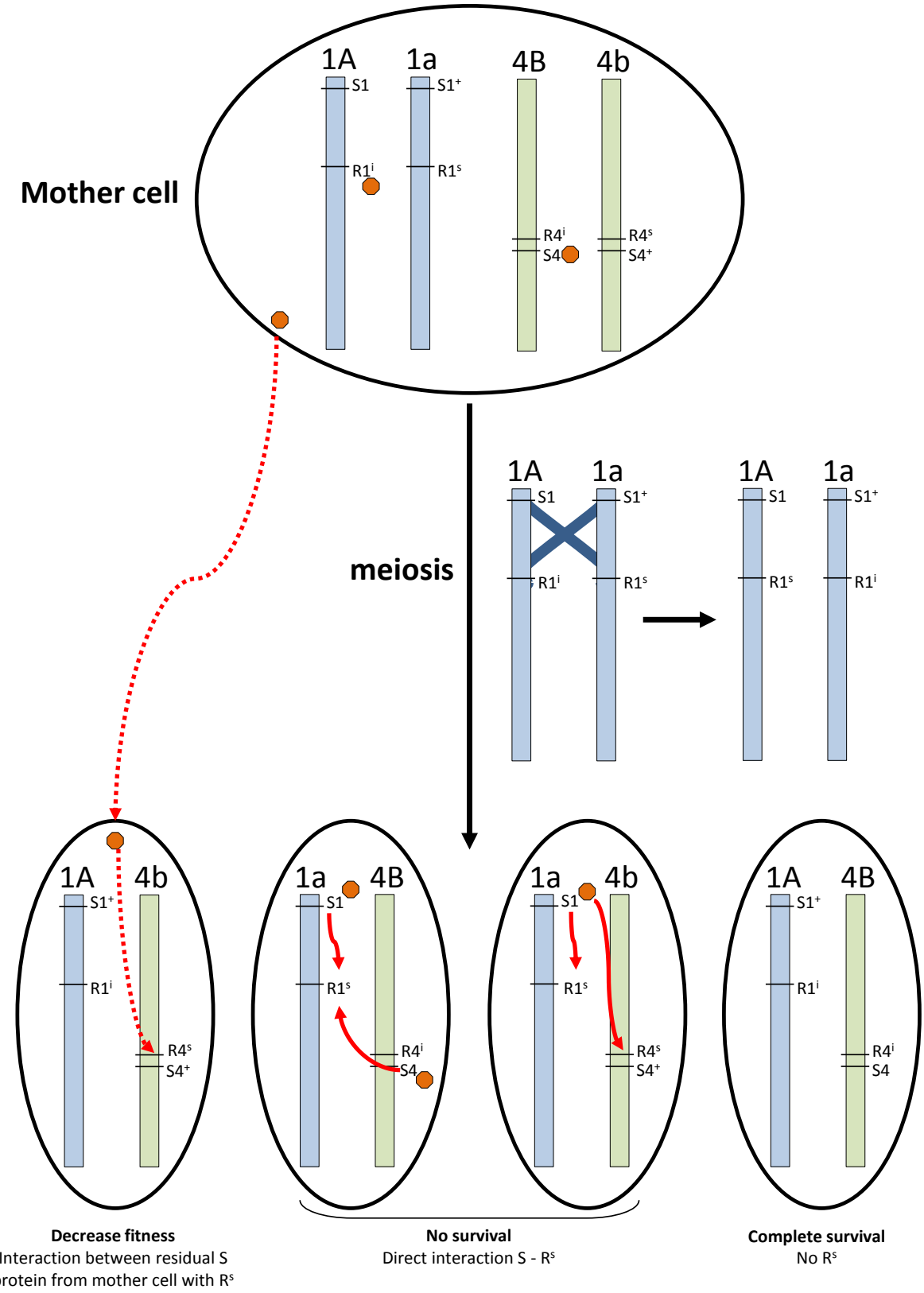


Figure 8: Hypothetical multiple loci interaction model with a recombination between S1 and R1 loci. The mother cell is heterozygous S1 $R1^{i}/S1^{+}$ $R1^{s}$ in chromosome 1 and S4 $R4^{i}/S4^{+}$ $R4^{s}$ in chromosome 4. The first gamete has its fitness reduced because of interaction with residual S proteins produced in mother cell and $R4^{s}$ allele. In gamete, the interaction of S alleles product with R^{s} allele is deleterious resulting in death of two gamete types. The last gamete has $R1^{i}$ and $R4^{i}$ allele and is thus not sensitive to S protein.

gametes (Figure 6). In this model, segregation distorter genes are noted S1/S1^{+} and S4/S4^{+} and responder are identified R1^{i}/R1^{s} and R4^{i}/R4^{s} for chromosomes 1 and 4 respectively. The product of S genes has a deleterious impact when it is in presence of R^{s} in gamete. Among the four produced gametes, two contain one Sx allele and one Rx^{s} allele and do not survive. The third gamete does not have S1 and S4 alleles but are R1^{s} and R4^{s}. Residual S proteins produced in mother cell create an interaction with R1^{s} and R4^{s} alleles and reduce this gamete type fitness. The last gamete type has both S1 and S4 alleles but no R1^{s} and R4^{s} alleles. This last gamete is not impacted by S proteins, it produced and show complete survival. The low recombination observed in chromosome 1 arm can be explained by the positioning of S1 and R1 loci at each extremity of the non-recombining region because in case of recombination between these regions more than half gametes are eliminated (Figure 7).

In general, gamete viability is associated to either paternal or maternal viability, for example in spermatid formation (reviewed in Larracuente and Presgraves, 2012), pollen germination failure (Mizuta et al., 2010), pollen sterility (Long et al., 2008; Yamagata et al., 2010), embryo-sac sterility (Chen et al., 2008; Yang et al., 2012) or preferential transmission of chromosomes to egg cell (Buckler et al., 1999; Longley, 1945; Rhoades, 1942). However, the two gene/genetic factor selection models proposed in banana had similar effect on both maternal and paternal gamete viability, implying that genetic factors involved in these distortions were not specific to maternal or paternal gamete formation pathways.

Finally, from the four models tested, only the duplication model can be excluded. The two translocation models could not alone fit the data but could fit if we include additional gene selection/gamete fitness. On the other hand, the four loci interaction model proposed here is sufficient to explain the observed segregation distortions, chromosome 1 and chromosome 4 marker pseudo-linkages and absence of recombination in chromosome 1. This model does not explain, at least in its present form, the multivalents observed during the meiosis of ‘Pahang’. However, it is not known if these multivalents involved chromosomes 1 and 4.

Additional approaches such as chromosome painting (Lysak et al., 2006) or ‘Pahang’ large segment pair-end sequencing associated to discordant read detection will be needed to conclude on the possible implication of structural heterozygosity on the distorted segregation observed in ‘Pahang’.

Bibliography

Bakry, F. (2008). Zygotic embryo rescue in bananas. Fruits 63, 111–115.

Bikard, D., Patel, D., Le Metté, C., Giorgi, V., Camilleri, C., Bennett, M.J., and Loudet, O. (2009). Divergent Evolution of Duplicate Genes Leads to Genetic Incompatibilities Within A. thaliana. Science 323, 623–626.

Birchler, J.A., and Veitia, R.A. (2007). The Gene Balance Hypothesis: From Classical Genetics to Modern Genomics. Plant Cell Online 19, 395–402.

Birchler, J.A., Dawe, R.K., and Doebley, J.F. (2003). Marcus Rhoades, Preferential Segregation and Meiotic Drive. Genetics 164, 835–841.

Buckler, E.S., Phelps-Durr, T.L., Buckler, C.S.K., Dawe, R.K., Doebley, J.F., and Holtsford, T.P. (1999). Meiotic Drive of Chromosomal Knobs Reshaped the Maize Genome. Genetics 153, 415–426.

Chen, J., Ding, J., Ouyang, Y., Du, H., Yang, J., Cheng, K., Zhao, J., Qiu, S., Zhang, X., Yao, J., et al. (2008). A triallelic system of S5 is a major regulator of the reproductive barrier and compatibility of indica–japonica hybrids in rice. Proc. Natl. Acad. Sci. 105, 11436–11441.

Cruz, V.M. (2013). Molecular Genetic Characterization of Lesquerella New Industrial Crop Using DArTseq Markers. In Plant and Animal Genome XXI Conference, (Plant and Animal Genome).

Dessauw, D. (1987). Etude des facteurs de la stérilité du bananier (Musa spp.) et des relations cytotaxinomiques entre M. acuminata Colla et M. balbisiana Colla. Université de Paris-sud.

D'Hont, A., Denoeud, F., Aury, J.-M., Baurens, F.-C., Carreel, F., Garsmeur, O., Noel, B., Bocs, S., Droc, G., Rouard, M., et al. (2012). The banana (Musa acuminata) genome and the evolution of monocotyledonous plants. Nature 488, 213–217.

Dobzhansky, T. (1937). Genetics and the Origin of Species (Columbia University Press).

Dobzhansky, T., and Sturtevant, A.H. (1938). Inversions in the chromosomes of Drosophila pseudoobscura. Genetics 23, 28.

Dodds, K.S. (1943). Genetical and cytological studies of Musa. V. Certain edible diploids. J. Genet. 45, 113–138.

Dodds, K., and Simmonds, N. (1948). Sterility and parthenocarpy in diploid hybrids of musa. Heredity 2, 101–117.

Fauré, S., Bakry, F., and González de Leon, D. (1993a). Cytogenetic studies of diploid bananas. In Breeding Banana and Plantain for Resistance to Diseases and Pests, (CIRAD-FLHOR, Montpellier: Ganry J.), pp. 77–92.

Fauré, S., Noyer, J.L., Horry, J.P., Bakry, F., Lanaud, C., and Gońzalez de León, D. (1993b). A molecular marker-based linkage map of diploid bananas (Musa acuminata). Theor. Appl. Genet. 87, 517–526.

Fishman, L., Kelly, A.J., Morgan, E., and Willis, J.H. (2001). A Genetic Map in the Mimulus guttatus Species Complex Reveals Transmission Ratio Distortion due to Heterospecific Interactions. Genetics 159, 1701–1716.

Hippolyte, I., Bakry, F., Seguin, M., Gardes, L., Rivallan, R., Risterucci, A.-M., Jenny, C., Perrier, X., Carreel, F., Argout, X., et al. (2010). A saturated SSR/DArT linkage map of Musa acuminata addressing genome rearrangements among bananas. BMC Plant Biol. 10, 65.

Jáuregui, B., de Vicente, M.C., Messeguer, R., Felipe, A., Bonnet, A., Salesses, G., and Arús, P. (2001). A reciprocal translocation between 'Garfi' almond and 'Nemared' peach. Theor. Appl. Genet. 102, 1169–1176.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–359.

Larracuente, A.M., and Presgraves, D.C. (2012). The Selfish Segregation Distorter Gene Complex of Drosophila melanogaster. Genetics 192, 33–53.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map (SAM) Format and SAMtools. Bioinformatics.

Long, Y., Zhao, L., Niu, B., Su, J., Wu, H., Chen, Y., Zhang, Q., Guo, J., Zhuang, C., Mei, M., et al. (2008). Hybrid male sterility in rice controlled by interaction between divergent alleles of two adjacent genes. Proc. Natl. Acad. Sci. 105, 18871–18876.

Longley, A.E. (1945). Abnormal segregation during megasporogenesis in maize. Genetics 30, 100.

Lysak, M.A., Berr, A., Pecinka, A., Schmidt, R., McBreen, K., and Schubert, I. (2006). Mechanisms of chromosome number reduction in Arabidopsis thaliana and related Brassicaceae species. Proc. Natl. Acad. Sci. U. S. A. 103, 5224–5229.

Martin, G., Baurens, F.-C., Droc, G., Rouard, M., Kilian, A., Hastie, A., Carreel, F., and D'Hont, A. (in prep.). A protocol to go beyond draft genome assemblies in plants: the banana sequence as a case study.

Mayr, E. (1942). Systematics and the origin of species, from the viewpoint of a zoologist (Harvard University Press).

Mbanjo, E., Tchoumbougnang, F., Mouelle, A., Oben, J., Nyine, M., Dochez, C., Ferguson, M., and Lorenzen, J. (2012). Molecular marker-based genetic linkage map of a diploid banana population (Musa acuminata Colla). Euphytica 188, 369–386.

Mizuta, Y., Harushima, Y., and Kurata, N. (2010). Rice pollen hybrid incompatibility caused by reciprocal gene loss of duplicated genes. Proc. Natl. Acad. Sci. 107, 20417–20422.

Noyer, J.L., Dambier, D., Lanaud, C., and Lagoda, P. (1997). The saturated map of diploid banana (Musa acuminata). In Abstract Plant & Animal Genome V Conference,.

Van Ooijen, J.W. (2011). Multipoint maximum likelihood mapping in a full-sib family of an outbreeding species. Genet. Res. 93, 343–349.

Ouyang, Y., and Zhang, Q. (2013). Understanding reproductive isolation based on the rice model. Annu. Rev. Plant Biol. 64, 111–135.

Perrier, X., and Jacquemoud-collet, J.-P. (2006). DARwin software.

Perrier, X., De Langhe, E., Donohue, M., Lentfer, C., Vrydaghs, L., Bakry, F., Carreel, F., Hippolyte, I., Horry, J.-P., Jenny, C., et al. (2011). Multidisciplinary perspectives on banana (Musa spp.) domestication. Proc. Natl. Acad. Sci. 108, 11311–11318.

Ramsey, J., Bradshaw, H.D., and Schemske, D.W. (2003). Components of reproductive isolation between the monkeyflowers Mimulus lewisii and M. cardinalis (Phrymaceae). Evolution 57, 1520–1534.

Rhoades, M.M. (1942). Preferential segregation in maize. Genetics 27, 395.

Rieseberg, L.H. (2001). Chromosomal rearrangements and speciation. Trends Ecol. Evol. 16, 351–358.

Risterucci, A.M., Grivet, L., N'Goran, J.A.K., Pieretti, I., Flament, M.H., and Lanaud, C. (2000). A high-density linkage map of Theobroma cacao L. Theor. Appl. Genet. 101, 948–955.

Shepherd, K. (1999). Cytogenetics of the genus Musa (IPGRI).

Spindel, J., Wright, M., Chen, C., Cobb, J., Gage, J., Harrington, S., Lorieux, M., Ahmadi, N., and McCouch, S. (2013). Bridging the genotyping gap: using genotyping by sequencing (GBS) to add high-density SNP markers and new value to traditional bi-parental mapping and breeding populations. Theor. Appl. Genet. 1–18.

Sweigart, A.L., and Willis, J.H. (2012). Molecular evolution and genetics of postzygotic reproductive isolation in plants. F1000 Biol. Rep. 4.

Takumi, S., Motomura, Y., Iehisa, J., and Kobayashi, F. (2013). Segregation distortion caused by weak hybrid necrosis in recombinant inbred lines of common wheat. Genetica 141, 463–470.

Vilarinhos, A.D. (2004). Cartographie génétique et cytogénétique chez le bananier: caractérisation des translocations.

White, F. (1962). Geographic variation and speciation in Africa with particular reference to Diospyros. In Taxonomy and Geography, (London: Nichols D), pp. 71–103.

Yamagata, Y., Yamamoto, E., Aya, K., Win, K.T., Doi, K., Sobrizal, Ito, T., Kanamori, H., Wu, J., Matsumoto, T., et al. (2010). Mitochondrial gene in the nuclear genome induces reproductive barrier in rice. Proc. Natl. Acad. Sci. 107, 1494–1499.

Yang, J., Zhao, X., Cheng, K., Du, H., Ouyang, Y., Chen, J., Qiu, S., Huang, J., Jiang, Y., Jiang, L., et al. (2012). A Killer-Protector System Regulates Both Hybrid Sterility and Segregation Distortion in Rice. Science 337, 1336–1340.

Supplementary Figures

This section contains 24 supplemental figures numbered from S1 to S24.

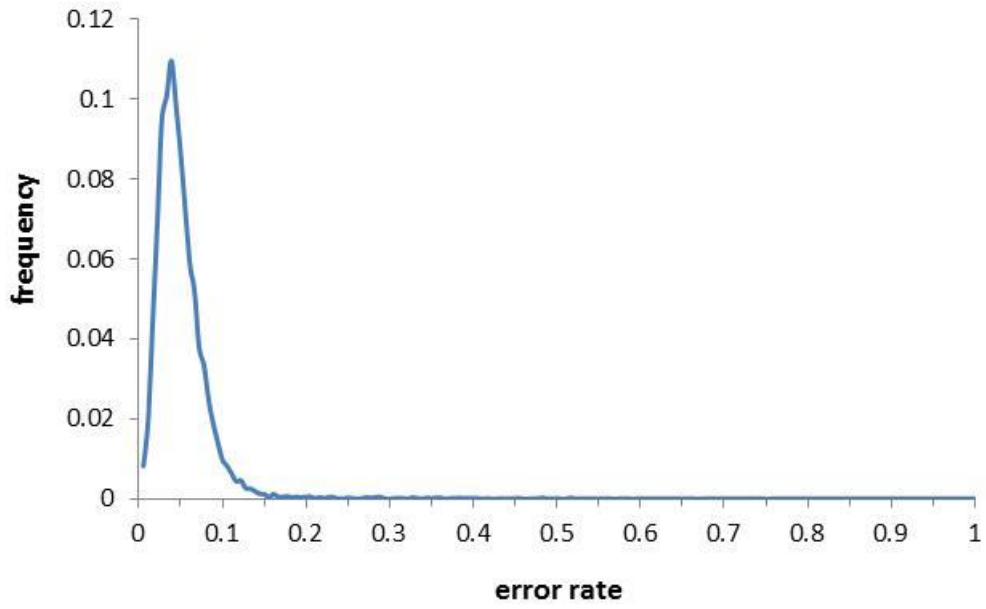
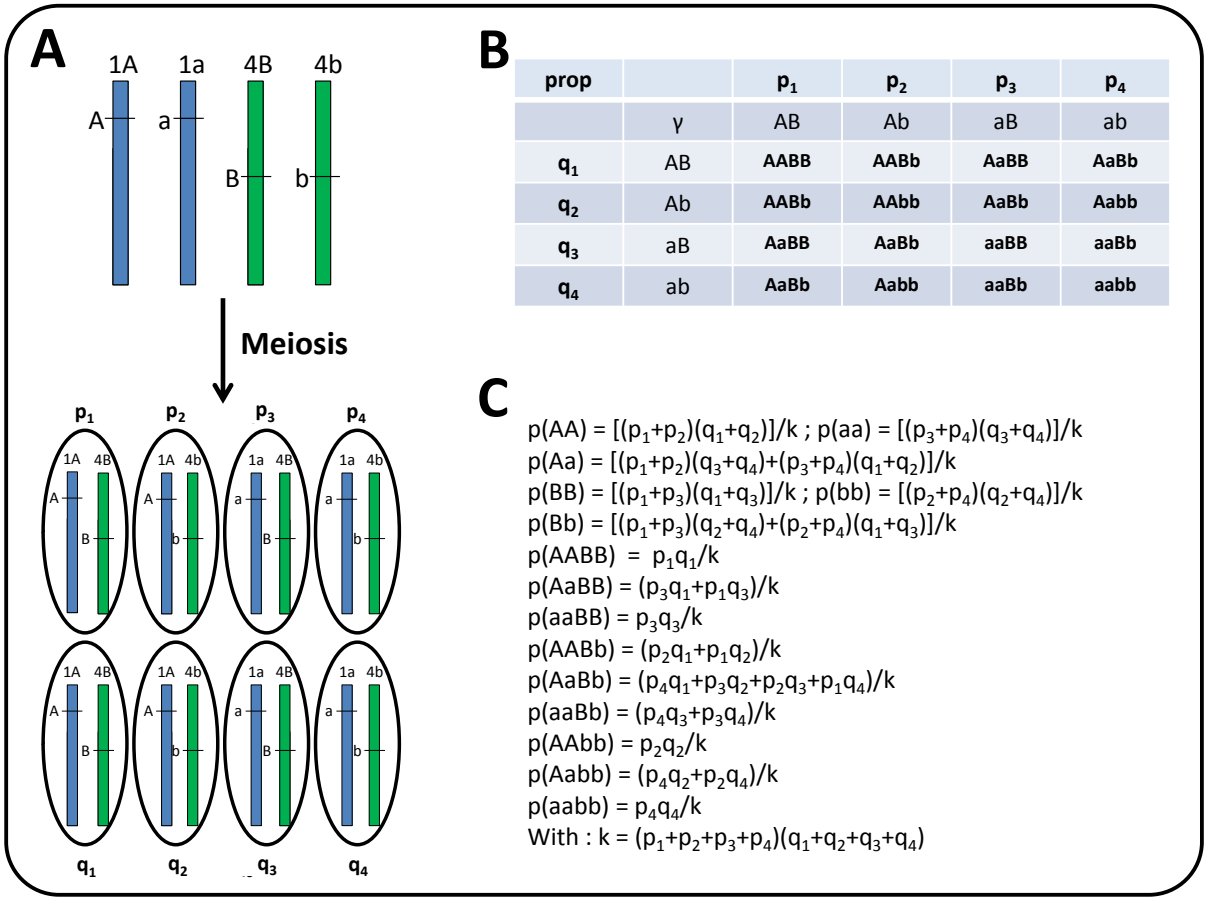


Figure S1: Distribution of the estimated genotyping error rate in the 7,417 co-dominant markers. An error is detected in a marker if for one individual the marker genotype compared to the consensus genotype of the 6 surrounding markers is different.



B

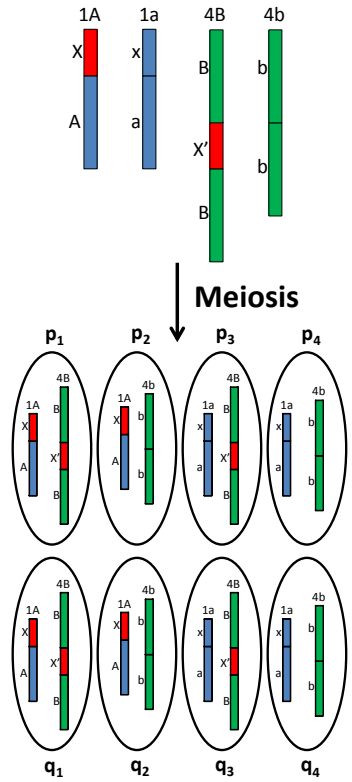
prop		p_1	p_2	p_3	p_4
	γ	AB	Ab	aB	ab
q_1	AB	**AABB**	**AABb**	**AaBB**	**AaBb**
q_2	Ab	**AABb**	**AAbb**	**AaBb**	**Aabb**
q_3	aB	**AaBB**	**AaBb**	**aaBB**	**aaBb**
q_4	ab	**AaBb**	**Aabb**	**aaBb**	**aabb**

C

$p(AA) = [(p_1+p_2)(q_1+q_2)]/k$; $p(aa) = [(p_3+p_4)(q_3+q_4)]/k$
$p(Aa) = [(p_1+p_2)(q_3+q_4)+(p_3+p_4)(q_1+q_2)]/k$
$p(BB) = [(p_1+p_3)(q_1+q_3)]/k$; $p(bb) = [(p_2+p_4)(q_2+q_4)]/k$
$p(Bb) = [(p_1+p_3)(q_2+q_4)+(p_2+p_4)(q_1+q_3)]/k$
$p(AABB) = p_1q_1/k$
$p(AaBB) = (p_3q_1+p_1q_3)/k$
$p(aaBB) = p_3q_3/k$
$p(AABb) = (p_2q_1+p_1q_2)/k$
$p(AaBb) = (p_4q_1+p_3q_2+p_2q_3+p_1q_4)/k$
$p(aaBb) = (p_4q_3+p_3q_4)/k$
$p(AAbb) = p_2q_2/k$
$p(Aabb) = (p_4q_2+p_2q_4)/k$
$p(aabb) = p_4q_4/k$
With : $k = (p_1+p_2+p_3+p_4)(q_1+q_2+q_3+q_4)$

Figure S2: (A) Schematic representation of gametes obtained from two pairs of chromosomes with **no structural heterozygosity.** A/a and B/b letters represent heterozygous loci on chromosomes 1 and 4. **(B)** Board of self-crossing of the gametes present in (A). p_1, p_2, p_3, p_4 and q_1, q_2, q_3, q_4 represent the survival probability of gametes from paternal and maternal parents respectively. **(C)** Equations predicting the genotype proportions as a function of (p_1, p_2, p_3, p_4, q_1, q_2, q_3, q_4) in case of no recombination.

A



B

prop		p_1	p_2	p_3	p_4
	γ	AB/XX'	Ab/X.	aB/xX'	ab/x.
q1	AB/XX'	**AABB/XXX'X'**	**AABb/XXX'.**	**AaBB/XxX'X'**	**AaBb/XxX'.**
q2	Ab/X.	**AABb/XXX'.**	**AAbb/XX..**	**AaBb/XxX'.**	**Aabb/Xx..**
q3	aB/xX'	**AaBB/XxX'X'**	**AaBb/XxX'.**	**aaBB/X'X'xx**	**aaBb/xxX'.**
q4	ab/x.	**AaBb/XxX'.**	**Aabb/Xx..**	**aaBb/xxX'.**	**aabb/xx..**

C

With X = X' (no divergence between X and X')

$p(AA) = [(p_1+p_2)(q_1+q_2)]/k$

$p(aa) = [(p_3+p_4)(q_3+q_4)]/k$

$p(Aa) = [(p_1+p_2)(q_3+q_4)+(p_3+p_4)(q_1+q_2)]/k$

$p(B) = [(p_1+p_3)(q_1+q_3)]/k$

$p(b) = [(p_2+p_4)(q_2+q_4)]/k$

$p(Bb) = [(p_1+p_3)(q_2+q_4)+(p_2+p_4)(q_1+q_3)]/k$

$P(X) = [p_1(q_1+q_2)+p_2(q_1+q_2)]/k$

$p(x) = p_4q_4/k$

$p(Xx) = [p_1(q_3+q_4)+p_2(q_3+q_4)+p_3(q_1+q_2+q_3+q_4)+p_4(q_1+q_2+q_3)]/k$

$p(AABB) = p_1q_1/k$

$p(AaBB) = (p_3q_1+p_1q_3)/k$

$p(aaBB) = p_3q_3/k$

$p(AABb) = (p_2q_1+p_1q_2)/k$

$p(AaBb) = (p_4q_1+p_3q_2+p_2q_3+p_1q_4)/k$

$p(aaBb) = (p_4q_3+p_3q_4)/k$

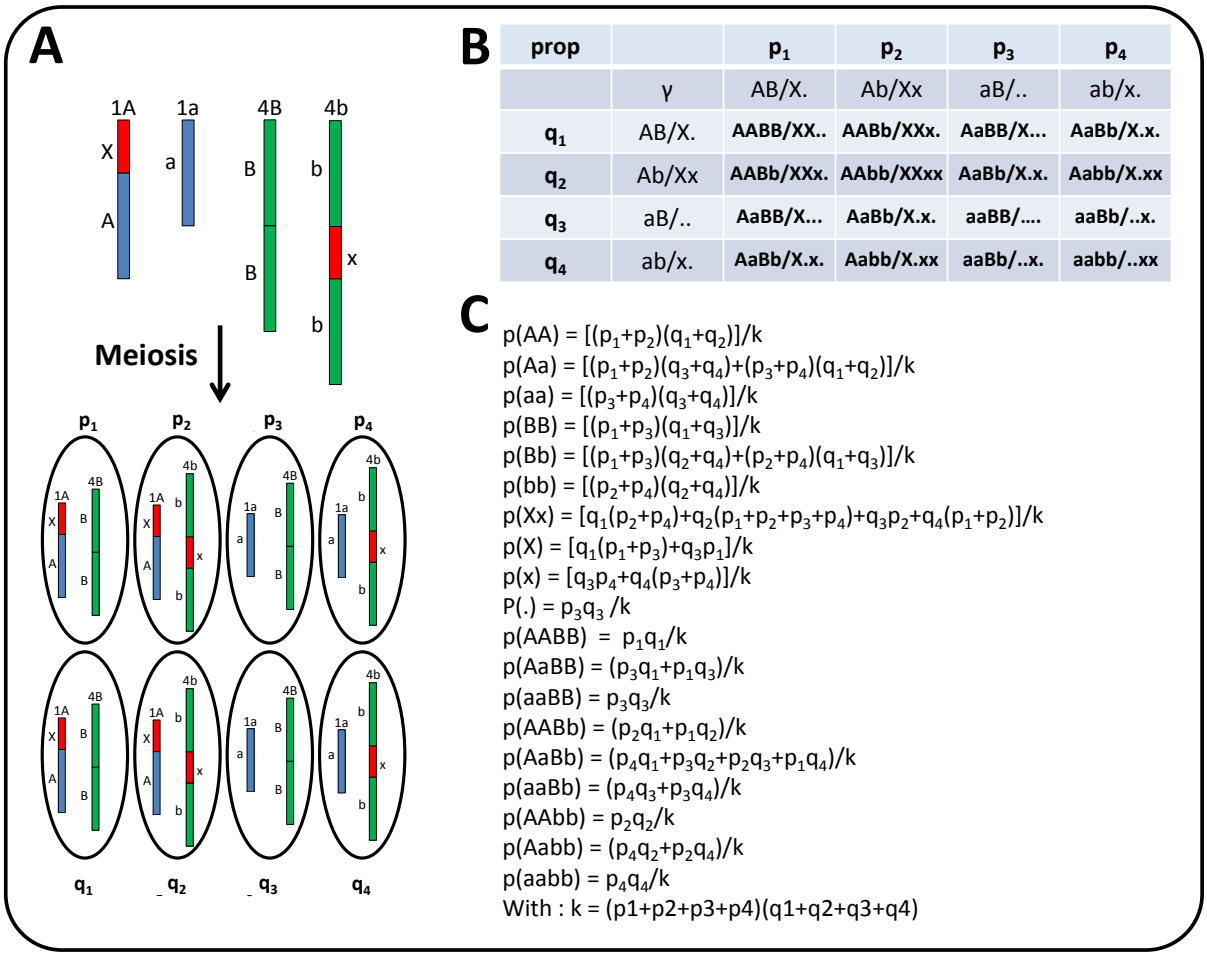
$p(AAbb) = p_2q_2/k$

$p(Aabb) = (p_4q_2+p_2q_4)/k$

$p(aabb) = p_4q_4/k$

With : k = (p1+p2+p3+p4)(q1+q2+q3+q4)

Figure S3: (A) Schematic representation of gametes obtained from two pairs of chromosomes with a **duplication** of a fragment of chromosome 1 in one of the homologs of chromosome 4. A/a, B/b, X/x/X' letters represent the distinct regions of chromosomes 1 and 4. **(B)** Board of self-crossing of the gametes present in (A). p_1, p_2, p_3, p_4 and q_1, q_2, q_3, q_4 represent the survival probability of gametes from paternal and maternal parents respectively. **(C)** Equations predicting the genotype proportions as a function of (p_1, p_2, p_3, p_4, q_1, q_2, q_3, q_4) in case of no recombination. For this case, the assumption of a recent duplication of X fragment (called X') into chromosome 4B is made. This assumption resulted in an absence of divergence between X and X' and then X=X'. P(X) stands for the probability to observe homozygous genotype X and contains XX, X., XXX., and XXXX genotypes. The same notation has been applied to P(x) and P(Xx).



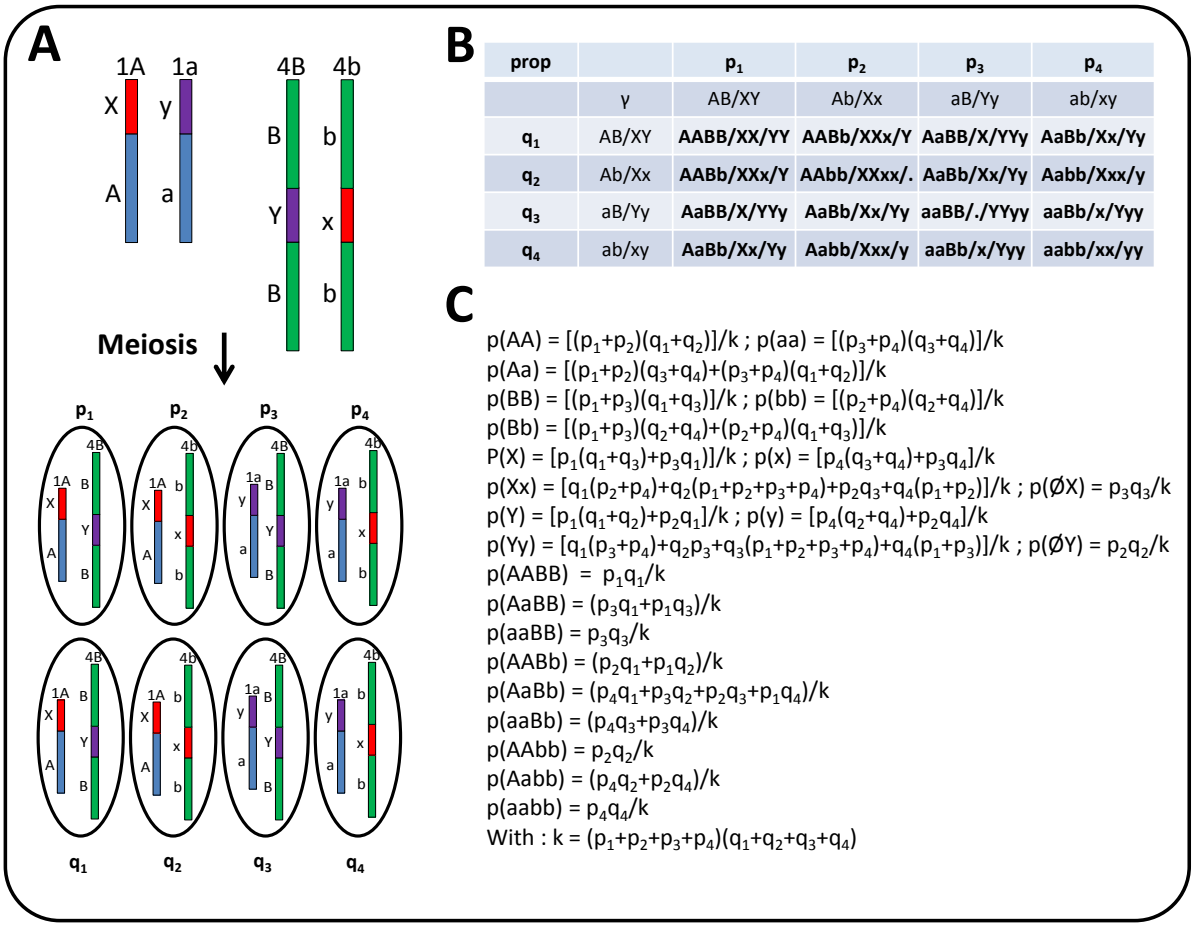
B

prop		p_1	p_2	p_3	p_4
	γ	AB/X.	Ab/Xx	aB/..	ab/x.
q_1	AB/X.	**AABB/XX..**	**AABb/XXx.**	**AaBB/X...**	**AaBb/X.x.**
q_2	Ab/Xx	**AABb/XXx.**	**AAbb/XXxx**	**AaBb/X.x.**	**Aabb/X.xx**
q_3	aB/..	**AaBB/X...**	**AaBb/X.x.**	**aaBB/....**	**aaBb/..x.**
q_4	ab/x.	**AaBb/X.x.**	**Aabb/X.xx**	**aaBb/..x.**	**aabb/..xx**

C

$p(AA) = [(p_1+p_2)(q_1+q_2)]/k$
$p(Aa) = [(p_1+p_2)(q_3+q_4)+(p_3+p_4)(q_1+q_2)]/k$
$p(aa) = [(p_3+p_4)(q_3+q_4)]/k$
$p(BB) = [(p_1+p_3)(q_1+q_3)]/k$
$p(Bb) = [(p_1+p_3)(q_2+q_4)+(p_2+p_4)(q_1+q_3)]/k$
$p(bb) = [(p_2+p_4)(q_2+q_4)]/k$
$p(Xx) = [q_1(p_2+p_4)+q_2(p_1+p_2+p_3+p_4)+q_3p_2+q_4(p_1+p_2)]/k$
$p(X) = [q_1(p_1+p_3)+q_3p_1]/k$
$p(x) = [q_3p_4+q_4(p_3+p_4)]/k$
$P(.) = p_3q_3/k$
$p(AABB) = p_1q_1/k$
$p(AaBB) = (p_3q_1+p_1q_3)/k$
$p(aaBB) = p_3q_3/k$
$p(AABb) = (p_2q_1+p_1q_2)/k$
$p(AaBb) = (p_4q_1+p_3q_2+p_2q_3+p_1q_4)/k$
$p(aaBb) = (p_4q_3+p_3q_4)/k$
$p(AAbb) = p_2q_2/k$
$p(Aabb) = (p_4q_2+p_2q_4)/k$
$p(aabb) = p_4q_4/k$
With : $k = (p1+p2+p3+p4)(q1+q2+q3+q4)$

Figure S4: (A) Schematic representation of gametes obtained from two pairs of chromosomes with a **translocation** of a fragment of chromosome 1 (X) in one of the homologs of chromosome 4 (x). A/a, B/b, X/x letters represent the distinct regions of chromosomes 1 and 4. **(B)** Board of self-crossing of the gametes present in (A). p_1, p_2, p_3, p_4 and q_1, q_2, q_3, q_4 represent the survival probability of gametes from paternal and maternal parents respectively. **(C)** Equations predicting the genotype proportions as a function of (p_1, p_2, p_3, p_4, q_1, q_2, q_3, q_4) in case of no recombination. P(X) stands for the probability to observe homozygous genotype X and contains X…, XX.., XXX., and XXXX genotypes. The same notation has been applied to P(x) and P(Xx). P(.) stands for genotype that has not the X/x region.



B

prop		p_1	p_2	p_3	p_4
	γ	AB/XY	Ab/Xx	aB/Yy	ab/xy
q_1	AB/XY	**AABB/XX/YY**	**AABb/XXx/Y**	**AaBB/X/YYy**	**AaBb/Xx/Yy**
q_2	Ab/Xx	**AABb/XXx/Y**	**AAbb/XXxx/.**	**AaBb/Xx/Yy**	**Aabb/Xxx/y**
q_3	aB/Yy	**AaBB/X/YYy**	**AaBb/Xx/Yy**	**aaBB/./YYyy**	**aaBb/x/Yyy**
q_4	ab/xy	**AaBb/Xx/Yy**	**Aabb/Xxx/y**	**aaBb/x/Yyy**	**aabb/xx/yy**

C

$p(AA) = [(p_1+p_2)(q_1+q_2)]/k$; $p(aa) = [(p_3+p_4)(q_3+q_4)]/k$
$p(Aa) = [(p_1+p_2)(q_3+q_4)+(p_3+p_4)(q_1+q_2)]/k$
$p(BB) = [(p_1+p_3)(q_1+q_3)]/k$; $p(bb) = [(p_2+p_4)(q_2+q_4)]/k$
$p(Bb) = [(p_1+p_3)(q_2+q_4)+(p_2+p_4)(q_1+q_3)]/k$
$P(X) = [p_1(q_1+q_3)+p_3q_1)]/k$; $p(x) = [p_4(q_3+q_4)+p_3q_4]/k$
$p(Xx) = [q_1(p_2+p_4)+q_2(p_1+p_2+p_3+p_4)+p_2q_3+q_4(p_1+p_2)]/k$; $p(ØX) = p_3q_3/k$
$p(Y) = [p_1(q_1+q_2)+p_2q_1]/k$; $p(y) = [p_4(q_2+q_4)+p_2q_4]/k$
$p(Yy) = [q_1(p_3+p_4)+q_2p_3+q_3(p_1+p_2+p_3+p_4)+q_4(p_1+p_3)]/k$; $p(ØY) = p_2q_2/k$
$p(AABB) = p_1q_1/k$
$p(AaBB) = (p_3q_1+p_1q_3)/k$
$p(aaBB) = p_3q_3/k$
$p(AABb) = (p_2q_1+p_1q_2)/k$
$p(AaBb) = (p_4q_1+p_3q_2+p_2q_3+p_1q_4)/k$
$p(aaBb) = (p_4q_3+p_3q_4)/k$
$p(AAbb) = p_2q_2/k$
$p(Aabb) = (p_4q_2+p_2q_4)/k$
$p(aabb) = p_4q_4/k$
With : $k = (p_1+p_2+p_3+p_4)(q_1+q_2+q_3+q_4)$

Figure S5: (A) Schematic representation of gametes obtained from two pairs of chromosomes with a **reciprocal translocation** of a fragment of chromosome 1 (X) with a fragment of chromosome 4 (Y). A/a, B/b, X/x and Y/x letters represent the distinct regions of chromosomes 1 and 4. **(B)** Board of self-crossing of the gametes present in (A). p_1, p_2, p_3, p_4 and q_1, q_2, q_3, q_4 represent the survival probability of gametes from paternal and maternal parents respectively. **(C)** Equations predicting the genotype proportions as a function of (p_1, p_2, p_3, p_4, q_1, q_2, q_3, q_4) in case of no recombination. p(X) stands for the probability to observe homozygous genotype X and contains X, and XX genotypes. The same notation has been applied to p(x), p(Xx), p(Y), p(y) and p(Yy). p(ØX) and p(ØY) stand for genotypes that have not the X/x and Y/y regions, respectively.

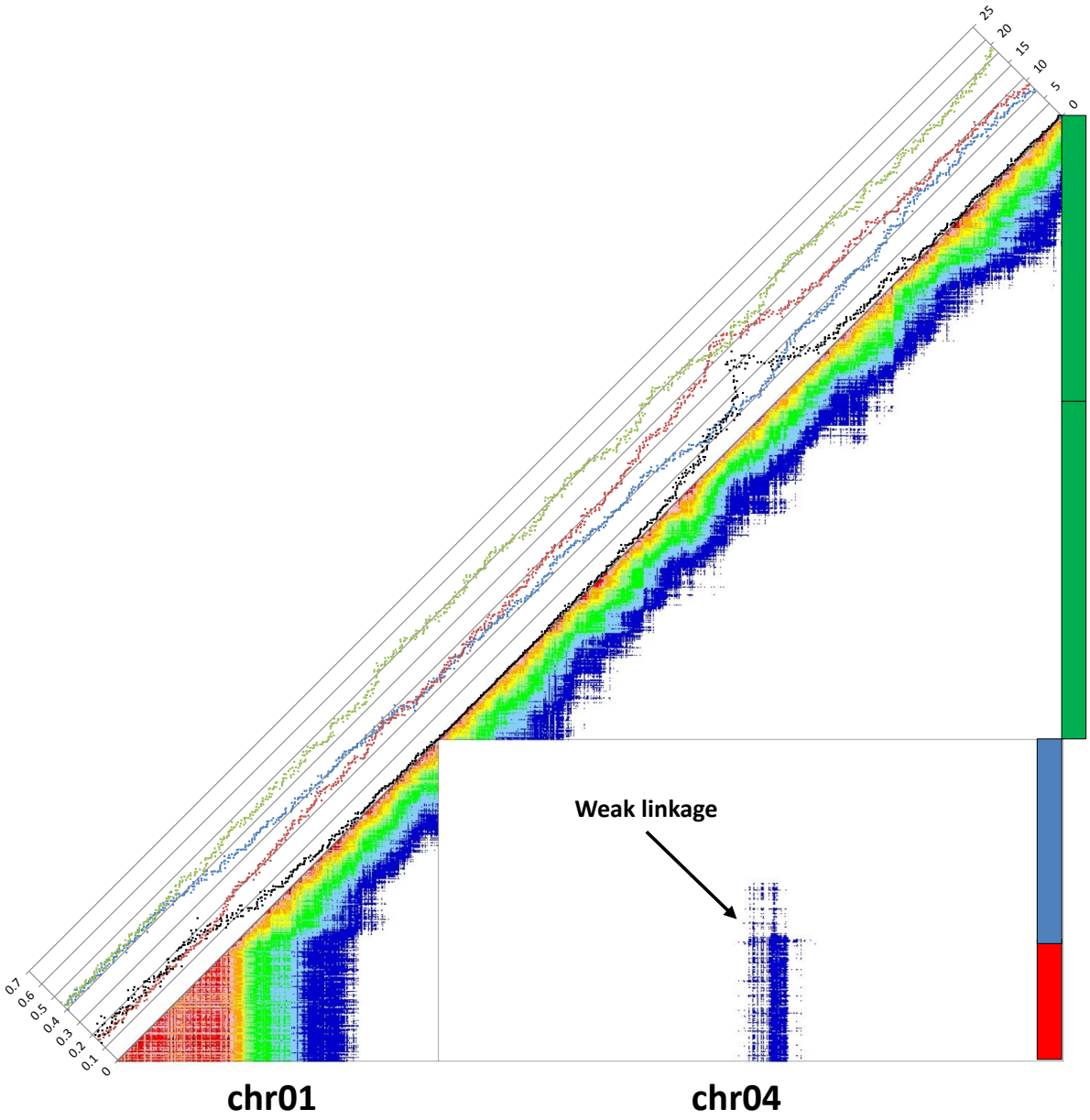


Figure S6: Marker linkage with co-dominant markers in case of duplication with gametic selection model proposed by Hippolyte et al., 2010. Dot-plot representing linkages between co-dominant markers of chromosomes 1 and 4 simulated on a self-progeny of 180 individuals. The simulated chromosomes of the parent of the self-progeny have a duplication of a fragment (red vertical block) of chromosome 1 (blue vertical block) in one of the homologs of chromosome 4 (green vertical block). No recombination is allowed in the duplicated fragment that corresponds to the first 120 markers of chromosome 1 of the standard structure described in Materials and Methods. The gametic selection proposed by Hippolyte et al 2010 has been applied on this simulated progeny. The linkage between markers is color-coded from strong linkages (warm colors) to no linkages (cold colors). The graph on the upper diagonal of the dot-plot represents for each marker the number of heterozygous, homozygous1 and homozygous2 individuals respectively in green, blue and red. The black dot graph represents the marker segregation distortions along the eleven chromosomes. The value presented in the graph is the $-\log_{10}$(p-value of the chi-square test testing the deviation from expected Mendelian segregation ratio).

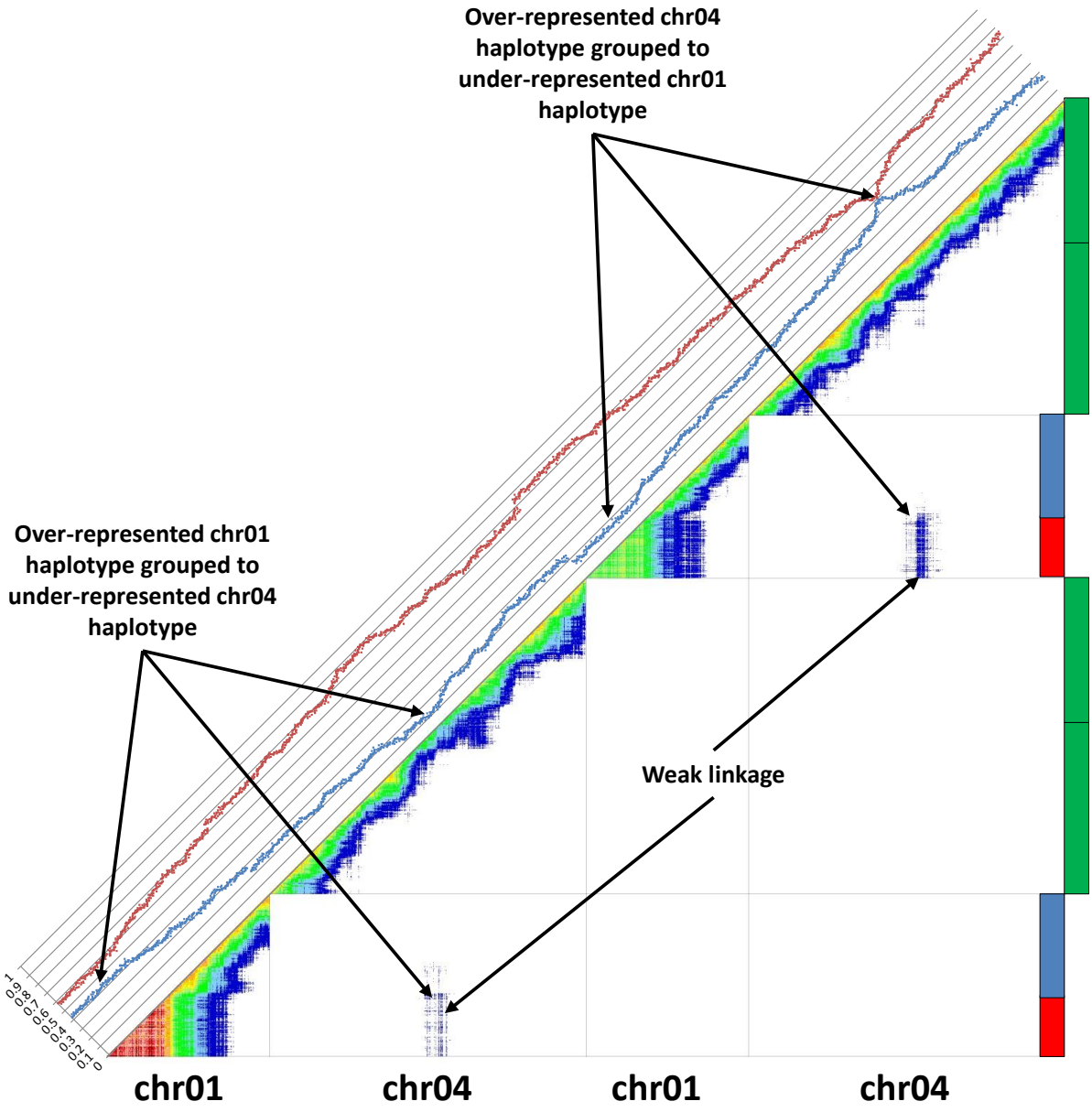


Figure S7: Marker linkage with dominant markers in case of duplication with gametic selection model proposed by Hippolyte et al., 2010. Dot-plot representing linkages between dominant markers of chromosomes 1 and 4 simulated on a self-progeny of 180 individuals. The simulated chromosomes of the parent of the self-progeny have a duplication of a fragment (red vertical block) of chromosome 1(blue vertical block) in one of the homologs of chromosome 4 (green vertical block). No recombination is allowed in the duplicated fragment that corresponds to the first 120 markers of chromosome 1 of the standard structure described in Materials and Methods. The gametic selection proposed by Hippolyte et al 2010 has been applied on this simulated progeny. The linkage between markers is color-coded from strong linkages (warm colors) to no linkages (cold colors). The graph on the upper diagonal of the dot-plot represents for each marker the proportion of recessive and dominant individuals respectively in blue and red.

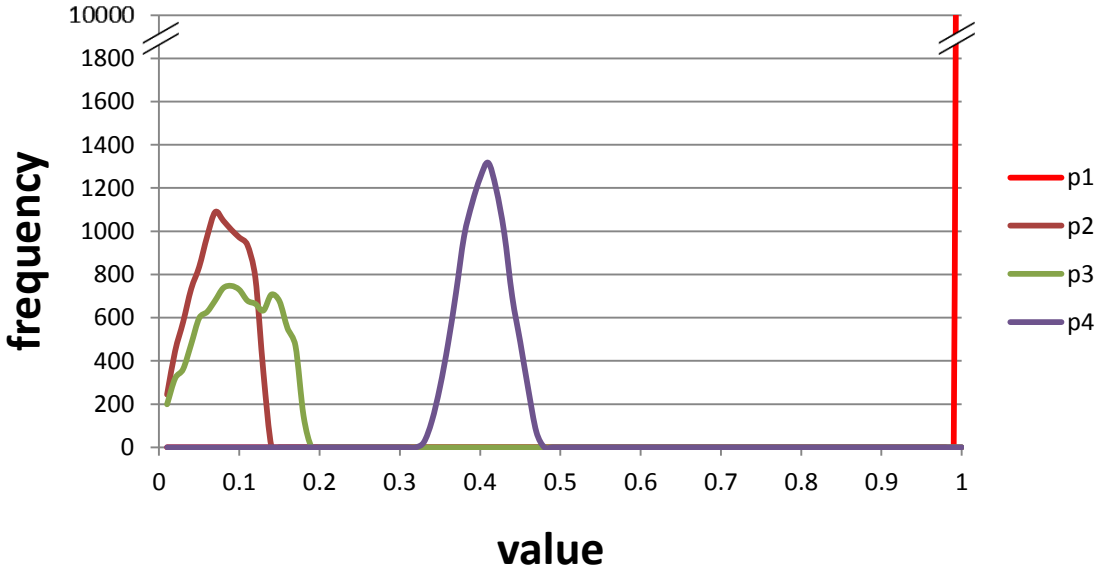


Figure S8: Plot showing the distribution of 10,000 possible values of p_1, p_2, p_3 and p_4 consistent with the observed genotype proportions in distorted regions and in case of no structural heterozygosity.

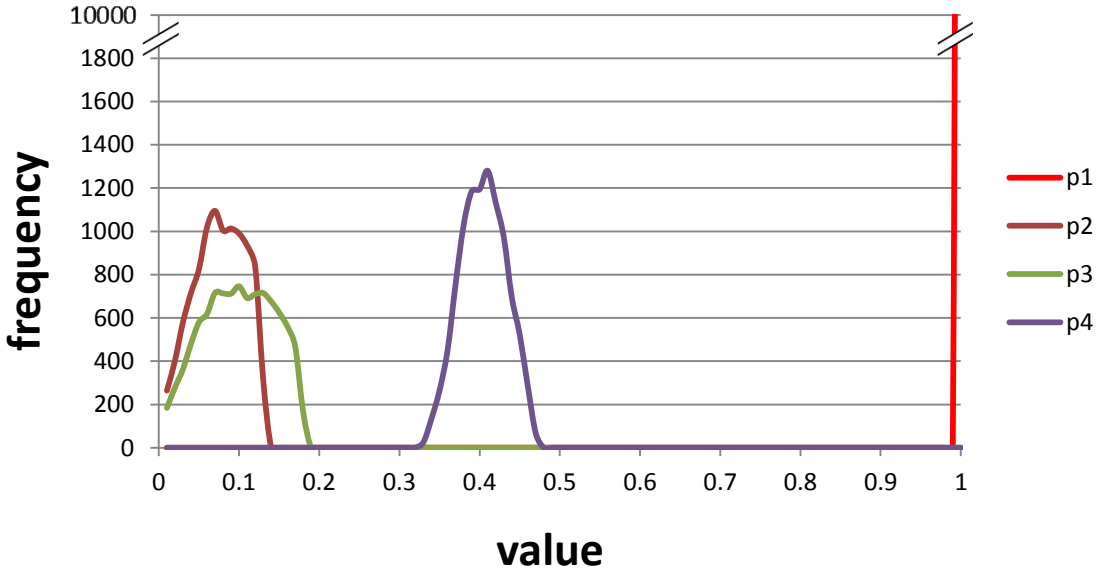


Figure S9: : Plot showing the distribution of 10,000 possible values of p_1, p_2, p_3 and p_4 consistent with the observed genotype proportions in distorted regions and in case of duplication.

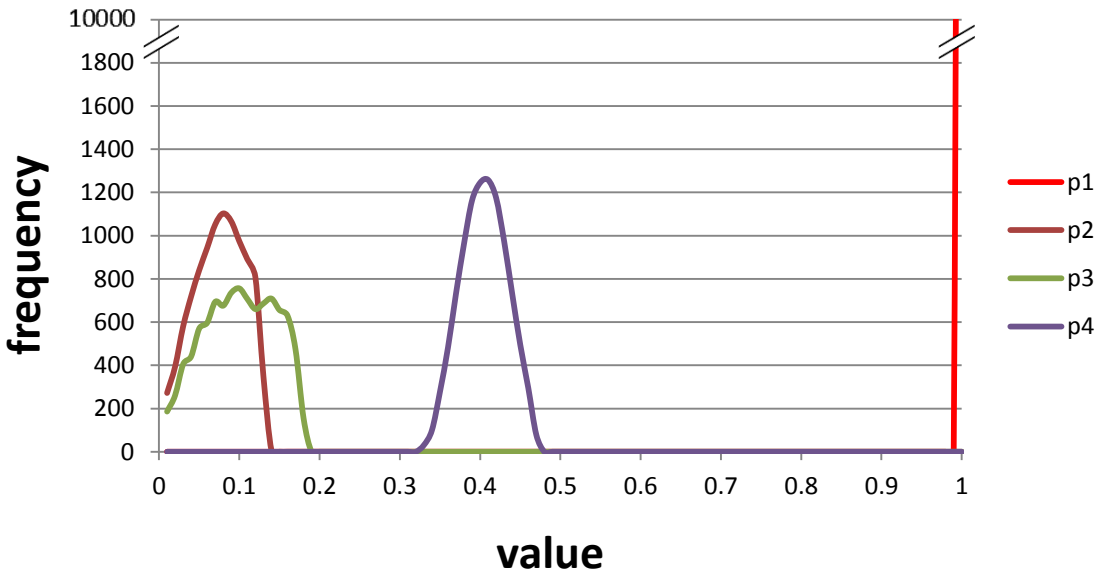


Figure S10: : Plot showing the distribution of 10,000 possible values of p_1, p_2, p_3 and p_4 consistent with the observed genotype proportions in distorted regions and in case of non-reciprocal translocation.

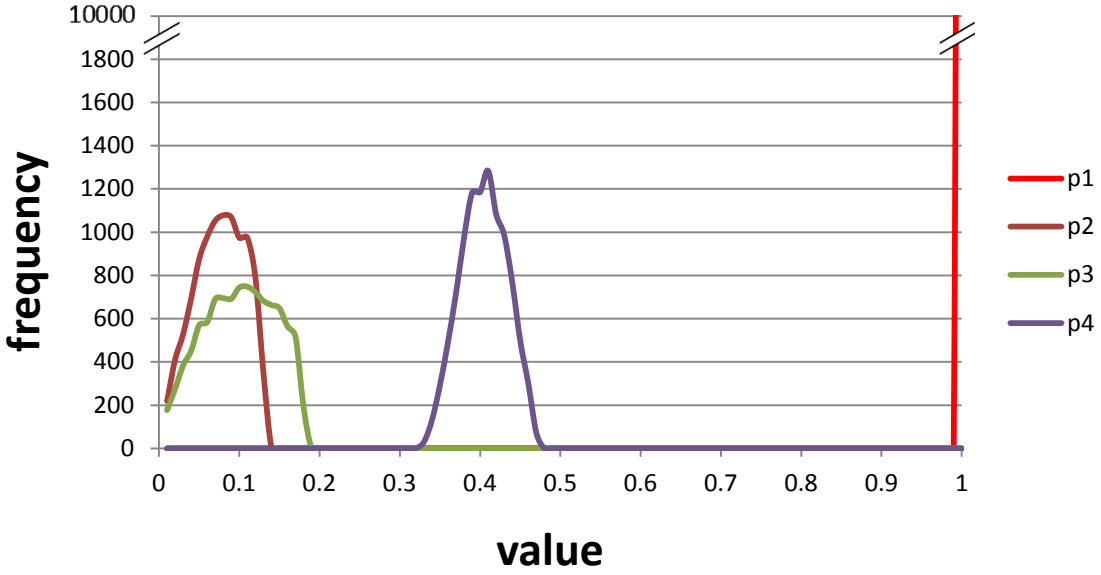


Figure S11: : Plot showing the distribution of 10,000 possible values of p_1, p_2, p_3 and p_4 consistent with the observed genotype proportions in distorted regions and in case of reciprocal translocation.

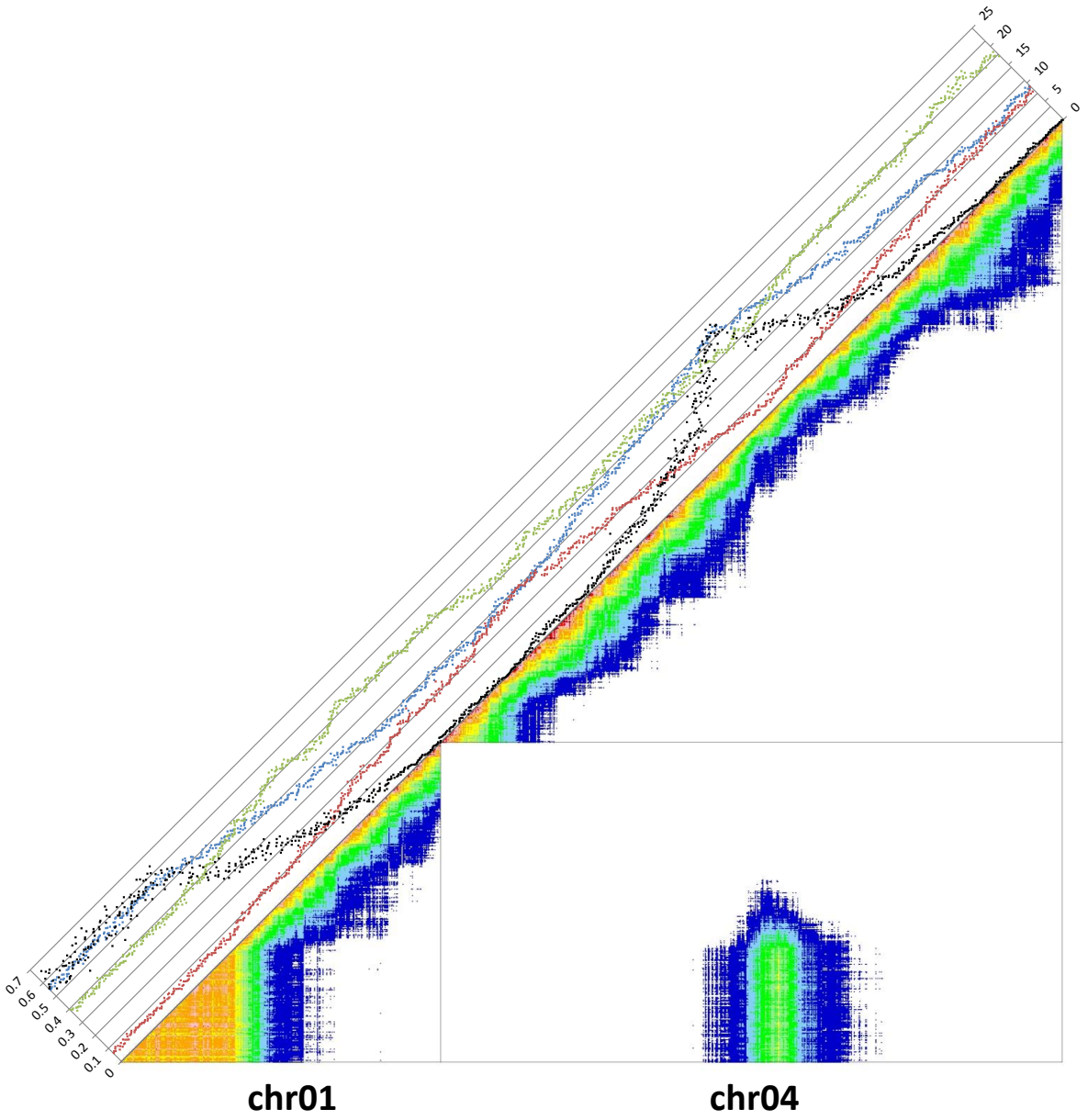


Figure S12: Marker linkage with co-dominant markers in case of no structural heterozygosity with gametic selection model. Dot-plot representing linkages between co-dominant markers of chromosomes 1 and 4 simulated on a self-progeny of 180 individuals. No recombination is allowed in the first 120 markers of chromosome 1. The gametic selection deduced in this paper $p_1 = 1$, $p_2 = 0.07$, $p_3 = 0.09$, $p_4 = 0.4$, $q_1 = 1$, $q_2 = 0$, $q_3 = 0$ and $q_4 = 0.28$ has been applied on this simulated progeny. The linkage between markers is color-coded from strong linkages (warm colors) to no linkages (cold colors). The graph on the upper diagonal of the dot-plot represents for each marker the number of heterozygous, homozygous1 and homozygous2 individuals respectively in green, blue and red. The black dot graph represents the marker segregation distortions along the eleven chromosomes. The value presented in the graph is the $-\log_{10}$(p-value of the chi-square test testing the deviation from expected Mendelian segregation ratio).

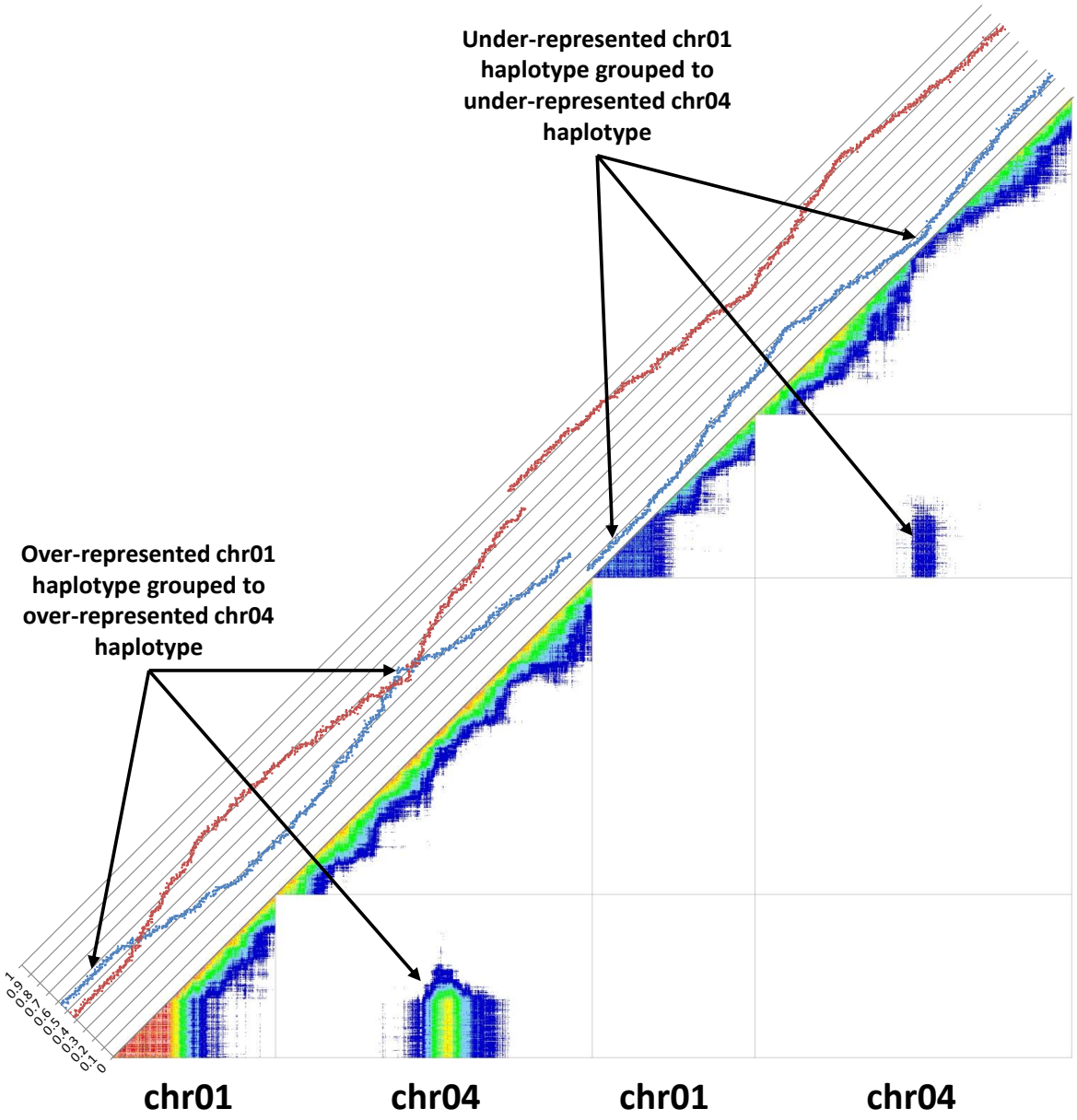


Figure S13: Marker linkage with dominant markers in case of no structural heterozygosity with gametic selection model. Dot-plot representing linkages between dominant markers of chromosomes 1 and 4 simulated on a self-progeny of 180 individuals. No recombination is allowed in the first 120 markers of chromosome 1. The gametic selection deduced in this paper ($p_1 = 1$, $p_2 = 0.07$, $p_3 = 0.09$, $p_4 = 0.4$, $q_1 = 1$, $q_2 = 0$, $q_3 = 0$ and $q_4 = 0.28$) has been applied on this simulated progeny. The linkage between markers is color-coded from strong linkages (warm colors) to no linkages (cold colors). The graph on the upper diagonal of the dot-plot represents for each marker the proportion of recessive and dominant individuals respectively in blue and red.

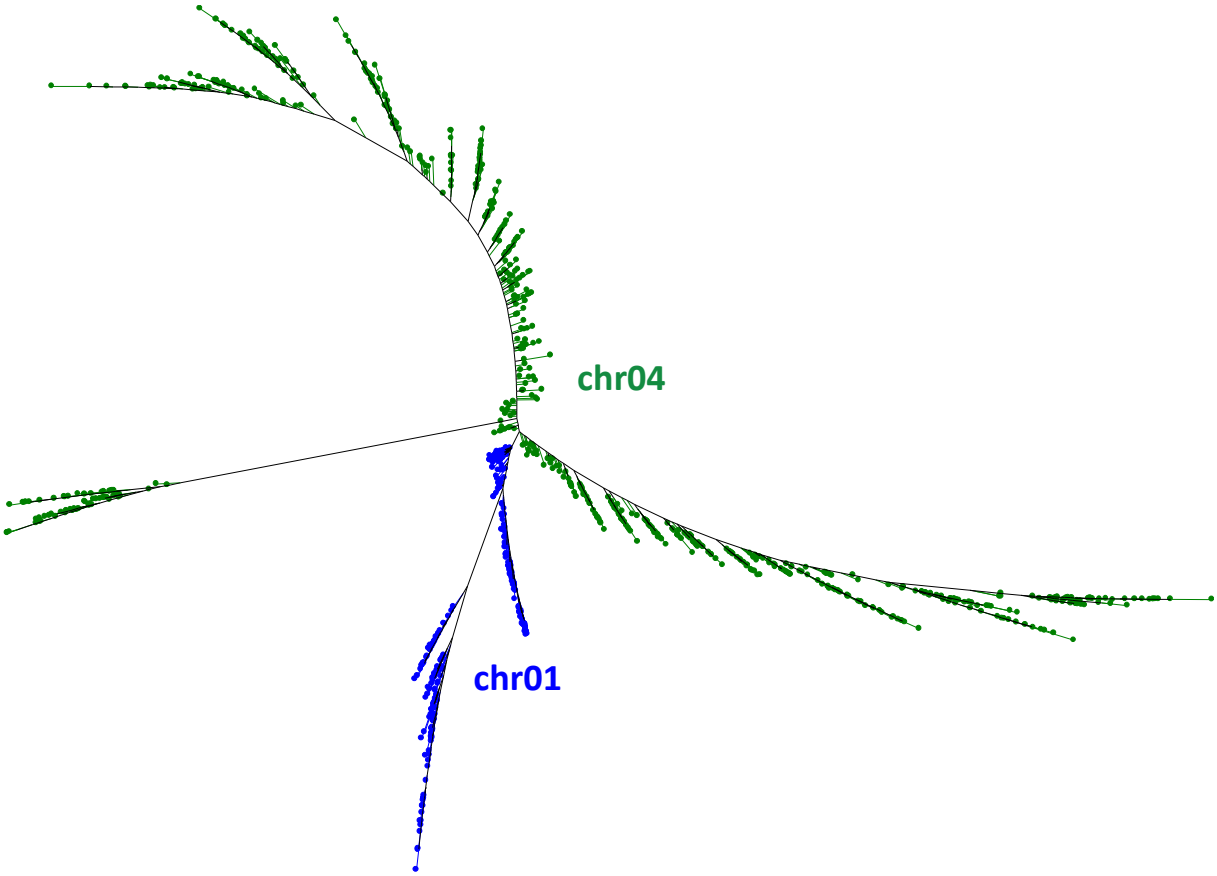


Figure S14: Weighted Neighbor-joining in case of no structural heterozygosity with gametic selection model. The tree has been constructed using the pairwise map distances calculated with the Kosambi mapping function on co-dominant markers of chromosomes 1 and 4, simulated on a self-progeny of 180 individuals. No recombination is allowed in the first 120 markers of chromosome 1. The gametic selection deduced in this paper ($p_1 = 1$, $p_2 = 0.07$, $p_3 = 0.09$, $p_4 = 0.4$, $q_1 = 1$, $q_2 = 0$, $q_3 = 0$ and $q_4 = 0.28$) has been applied on this simulated progeny. The distances were calculated from the pairwise recombination frequencies calculated by JoinMap4.1. Markers belonging to chromosome 1 and chromosome 4 are drawn in blue and green, respectively.

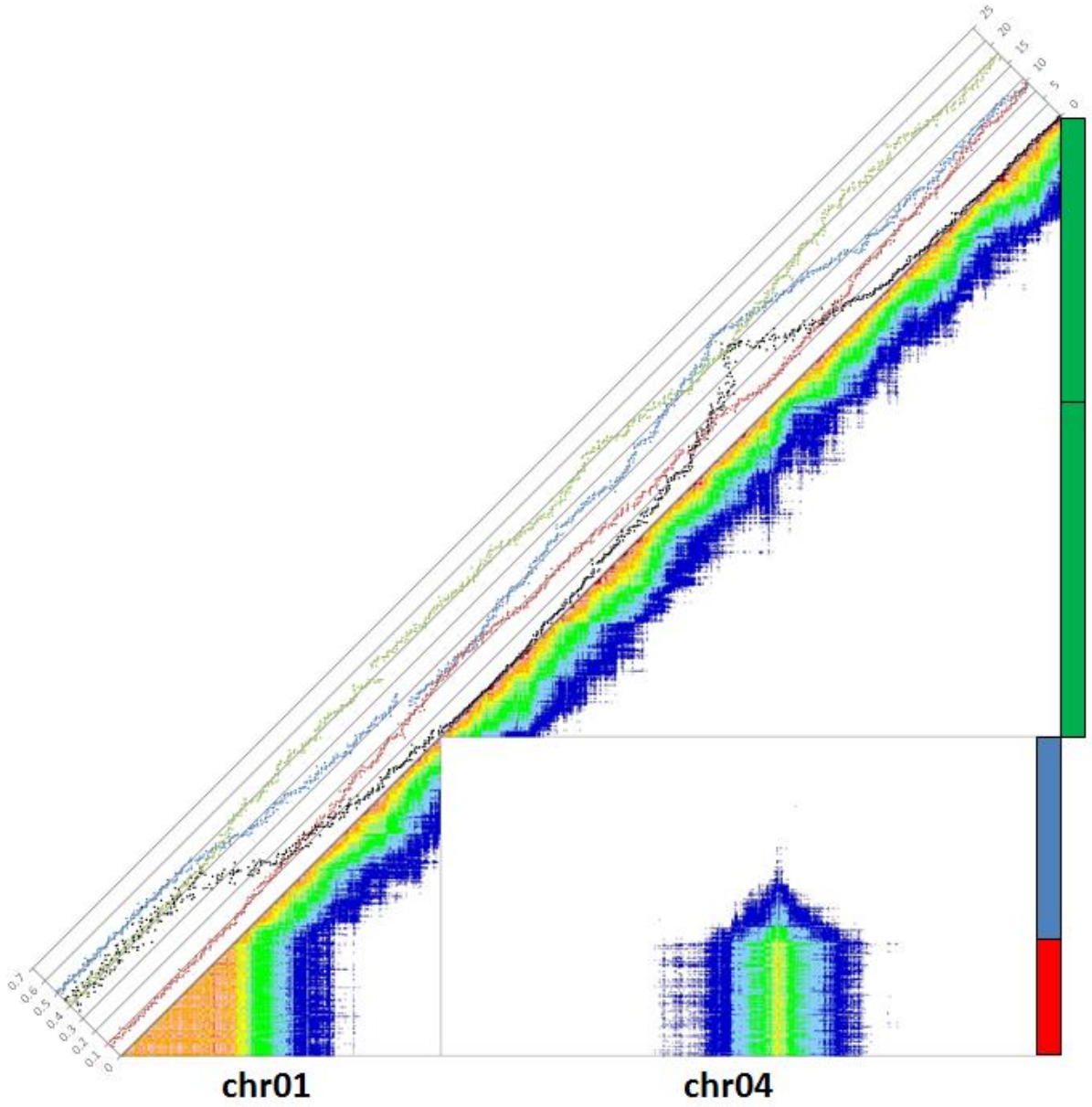


Figure S15: Marker linkage with co-dominant markers in case of duplication with gametic selection model. Dot-plot representing linkages between co-dominant markers of chromosomes 1 and 4 simulated on a self-progeny of 180 individuals. The simulated chromosomes of the parent of the self-progeny have a duplication of a fragment (red vertical block) of chromosome 1(blue vertical block) in one of the homologs of chromosome 4 (green vertical block). No recombination is allowed in the duplicated fragment that corresponds to the first 120 markers of chromosome 1 of the standard structure. The gametic selection deduced in this paper ($p_1 = 1$, $p_2 = 0.07$, $p_3 = 0.09$, $p_4 = 0.4$, $q_1 = 1$, $q_2 = 0$, $q_3 = 0$ and $q_4 = 0.28$) has been applied on this simulated progeny. The linkage between markers is color-coded from strong linkages (warm colors) to no linkages (cold colors). The graph on the upper diagonal of the dot-plot represents for each marker the number of heterozygous, homozygous1 and homozygous2 individuals respectively in green, blue and red. The black dot graph represents the marker segregation distortions along the eleven chromosomes. The value presented in the graph is the $-\log_{10}$(p-value of the chi-square test testing the deviation from expected Mendelian segregation ratio).

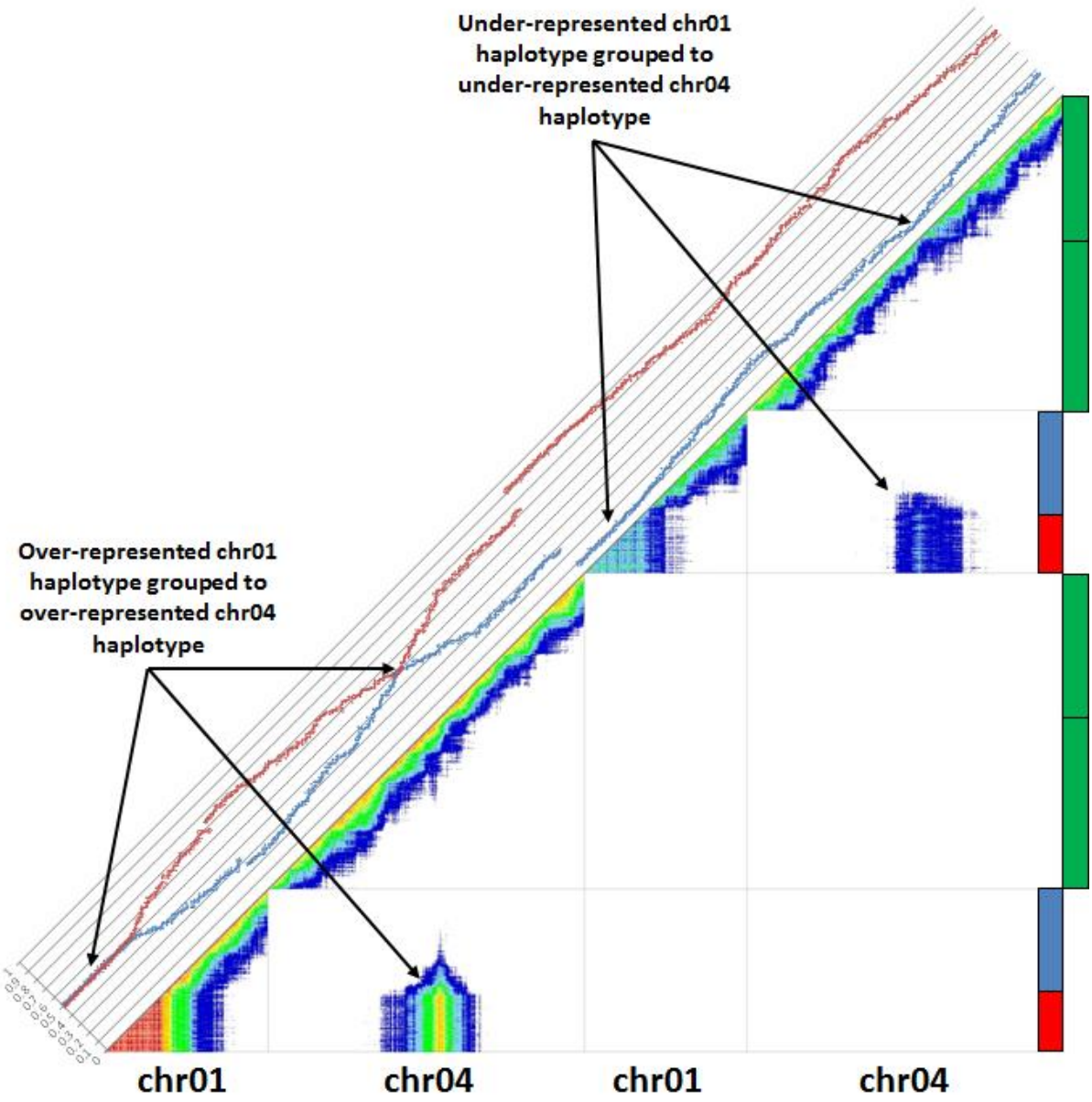


Figure S16: Marker linkage with co-dominant markers in case of duplication with gametic selection model. Dot-plot representing linkages between dominant markers of chromosomes 1 and 4 simulated on a self-progeny of 180 individuals. The simulated chromosomes of the parent of the self-progeny have a duplication of a fragment (red vertical block) of chromosome 1(blue vertical block) in one of the homologs of chromosome 4 (green vertical block). No recombination is allowed in the duplicated fragment that corresponds to the first 120 markers of chromosome 1 of the standard structure. The gametic selection deduced in this paper ($p_1 = 1$, $p_2 = 0.07$, $p_3 = 0.09$, $p_4 = 0.4$, $q_1 = 1$, $q_2 = 0$, $q_3 = 0$ and $q_4 = 0.28$) has been applied on this simulated progeny. The linkage between markers is color-coded from strong linkages (warm colors) to no linkages (cold colors). The graph on the upper diagonal of the dot-plot represents for each marker the proportion of recessive and dominant individuals respectively in blue and red.

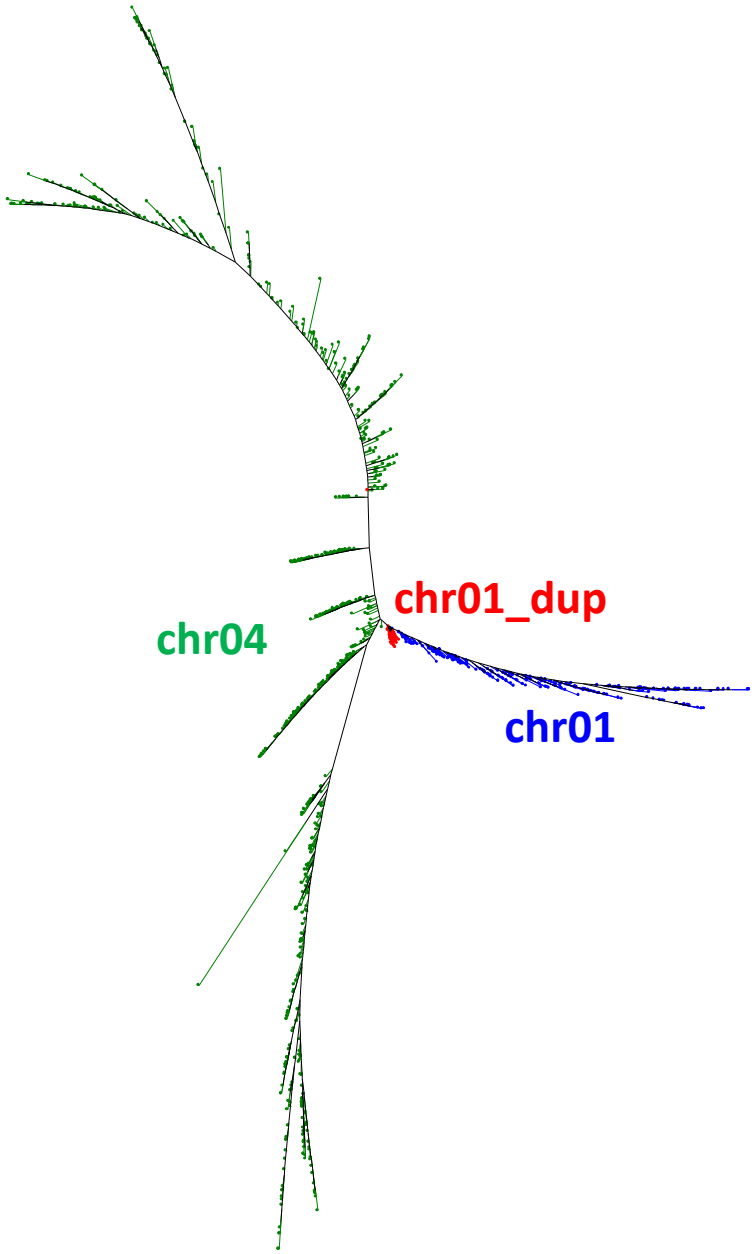


Figure S17: Weighted Neighbor-joining in case of duplication with gametic selection model. The tree has been constructed using the pairwise map distances calculated with the Kosambi mapping function on co-dominant markers of chromosomes 1 and 4, simulated on a self-progeny of 180 individuals. The simulated chromosomes of the parent of the self-progeny have a duplication of a fragment of chromosome 1 in one of the homologs of chromosome 4. No recombination is allowed in the duplicated fragment that corresponds to the first 120 markers of chromosome 1 of the standard structure. The gametic selection deduced in this paper ($p_1 = 1$, $p_2 = 0.07$, $p_3 = 0.09$, $p_4 = 0.4$, $q_1 = 1$, $q_2 = 0$, $q_3 = 0$ and $q_4 = 0.28$) has been applied on this simulated progeny. The distances were calculated from the pairwise recombination frequencies calculated by JoinMap4.1. Markers belonging to chromosome 1 and chromosome 4 are drawn in blue and green respectively. Markers belonging to the duplicated fragment are drawn in red.

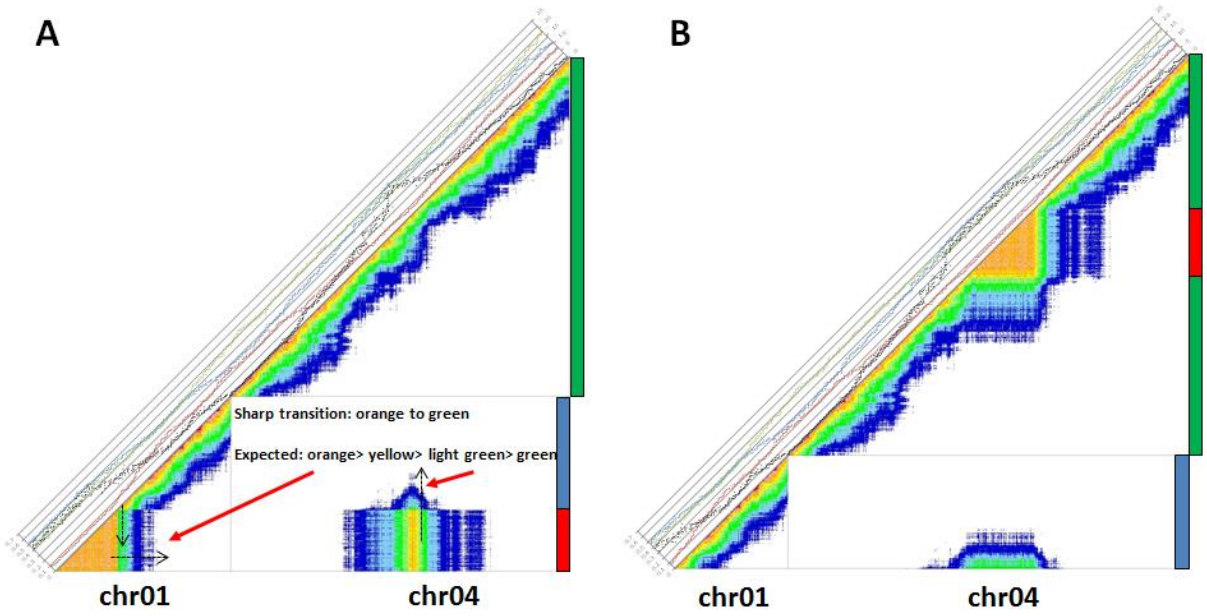


Figure S18: Marker linkage with co-dominant markers in case of non-reciprocal translocation with gametic selection model. Dot-plot representing linkages between co-dominant markers of chromosomes 1 and 4 simulated on a self-progeny of 180 individuals. The simulated chromosomes of the parent of the self-progeny have a translocation (red vertical block) of the beginning of chromosome 1 (blue vertical block) in one of the homologs of chromosome 4 (green vertical block). No recombination is allowed in the translocated fragment that corresponds to the first 120 markers of chromosome 1 in the standard structure. The gametic selection deduced in this paper ($p_1 = 1$, $p_2 = 0.07$, $p_3 = 0.09$, $p_4 = 0.4$, $q_1 = 1$, $q_2 = 0$, $q_3 = 0$ and $q_4 = 0.28$) has been applied on this simulated progeny. The linkage between markers is color coded from strong linkages (warm colors) to no linkages (cold colors). The graph on the upper diagonal of the dot-plot represents for each marker the number of heterozygous, homozygous1 and homozygous2 individuals respectively in green, blue and red. The black dot graph represents the marker segregation distortions along the eleven chromosomes. The value presented in the graph is the $-\log_{10}$(p-value of the chi-square test testing the deviation from expected Mendelian segregation ratio). (**A**) Dot plot with markers organized to the haplotype structure where the translocated fragment is at the extremity of chromosome 1. Breakpoints (red arrows) in marker linkage can be observed in the junction between markers located on translocated fragment and chromosome 1 (**B**) Dot plot with markers organized to the haplotype structure where the translocated fragment is in chromosome 4. No breakpoint is observed and there are few markers of chromosome 1 linked to markers of chromosome 4.

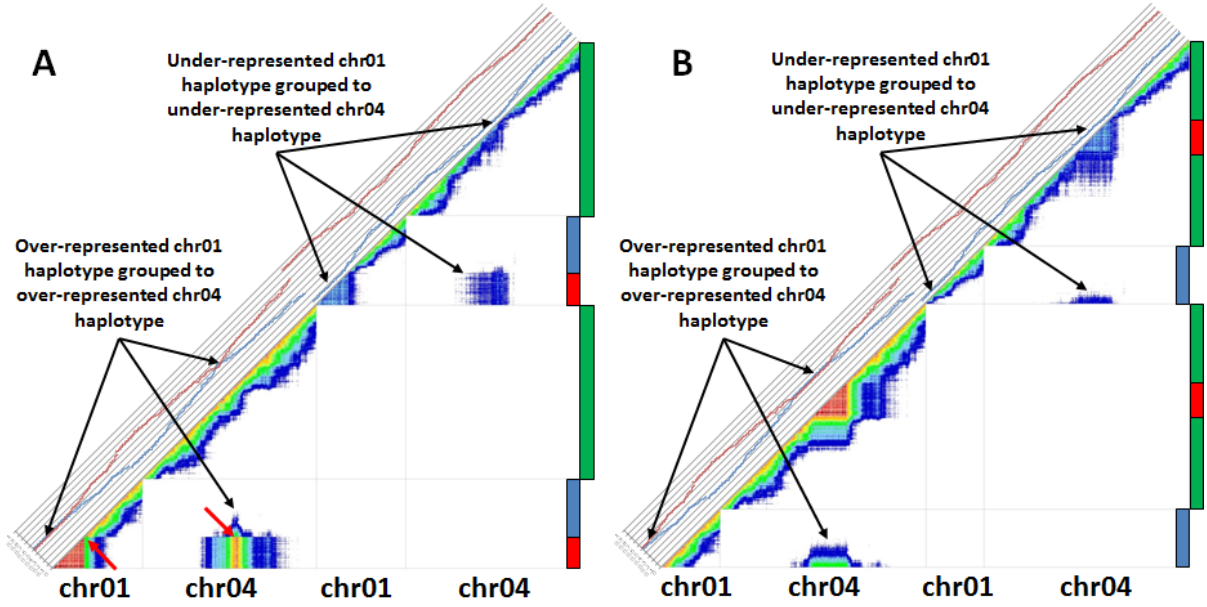


Figure S19: Marker linkage with dominant markers in case of non-reciprocal translocation with gametic selection model. Dot-plot representing linkages between dominant markers of chromosomes 1 and 4 simulated on a self-progeny of 180 individuals. The simulated chromosomes of the parent of the self-progeny have a translocation (red vertical block) of the beginning of chromosome 1 (blue vertical block) in one of the homologs of chromosome 4 (green vertical block). No recombination is allowed in the translocated fragment that corresponds to the first 120 markers of chromosome 1 in the standard structure. The gametic selection deduced in this paper ($p_1 = 1$, $p_2 = 0.07$, $p_3 = 0.09$, $p_4 = 0.4$, $q_1 = 1$, $q_2 = 0$, $q_3 = 0$ and $q_4 = 0.28$) has been applied on this simulated progeny. The linkage between markers is color-coded from strong linkages (warm colors) to no linkages (cold colors). The graph on the upper diagonal of the dot-plot represents for each marker the proportion of recessive and dominant individuals respectively in blue and red. (**A**) Dot-plot with markers organized to the haplotype structure where the translocated fragment is at the extremity of chromosome 1. For the over-represented genotype combinations, breakpoints (red arrows) in marker linkage can be observed in the junction between markers located on translocated fragment and chromosome 1 (**B**) Dot-plot with markers organized to the haplotype structure where the translocated fragment is in chromosome 4. No breakpoint is observed and there are few markers of chromosome 1 linked to markers of chromosome 4.

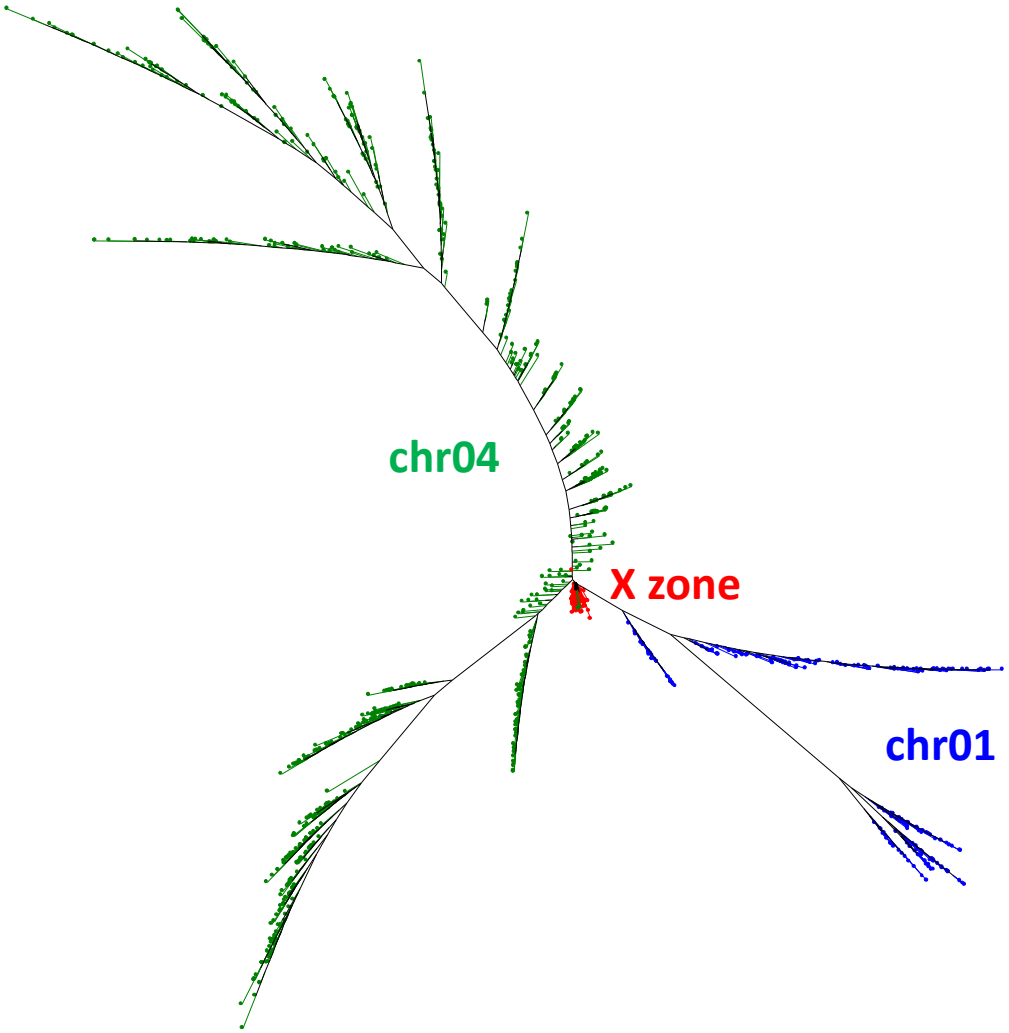


Figure S20: Weighted Neighbor-joining in case of non-reciprocal translocation with gametic selection model. The tree has been constructed using the pairwise map distances calculated with the Kosambi mapping function on co-dominant markers of chromosomes 1 and 4, simulated on a self-progeny of 180 individuals. The simulated chromosomes of the parent of the self-progeny have a translocation of the beginning of chromosome 1 in one of the homologs of chromosome 4. No recombination is allowed in the translocated fragment that corresponds to the first 120 markers of chromosome 1 in the standard structure. The gametic selection deduced in this paper ($p_1 = 1$, $p_2 = 0.07$, $p_3 = 0.09$, $p_4 = 0.4$, $q_1 = 1$, $q_2 = 0$, $q_3 = 0$ and $q_4 = 0.28$) has been applied on this simulated progeny. The distances were calculated from the pairwise recombination frequencies calculated by JoinMap4.1. Markers belonging to chromosome 1 and chromosome 4 are drawn in blue and green respectively. Markers belonging to the translocated fragment are drawn in red.

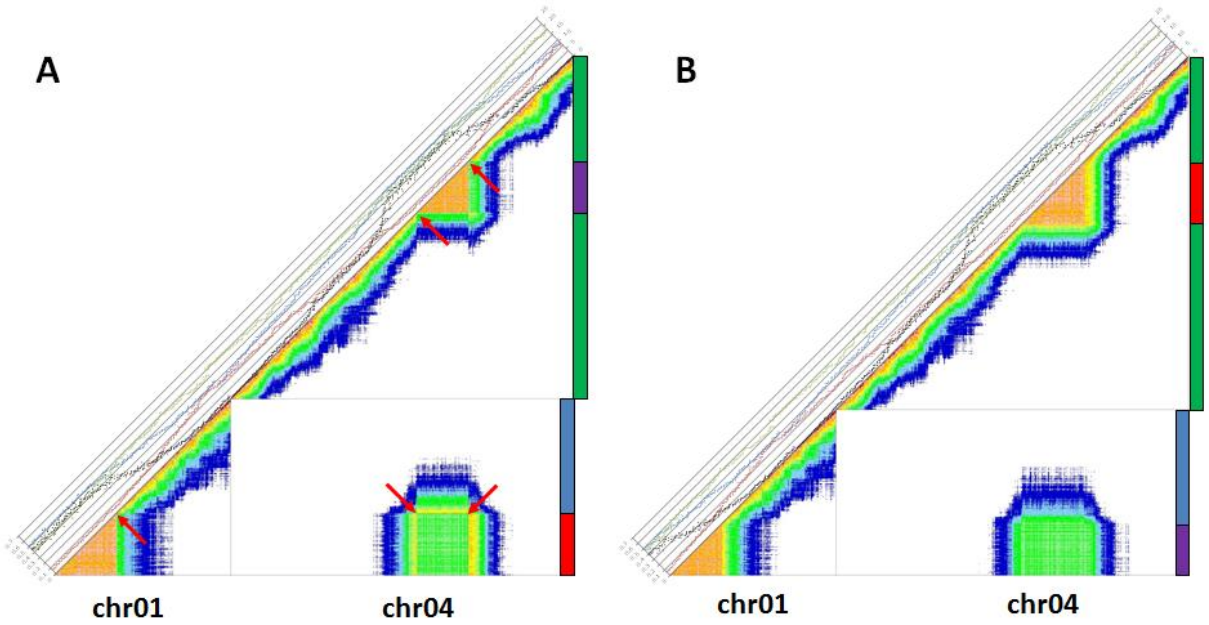


Figure S21: Marker linkage with co-dominant markers in case of reciprocal translocation with gametic selection model. Dot-plot representing linkages between co-dominant markers of chromosomes 1 and 4 simulated on a self-progeny of 180 individuals. The simulated chromosomes of the parent of the self-progeny have a reciprocal translocation (red and purple vertical blocks) of the beginning of chromosome 1 (blue vertical block) with a fragment of chromosome 4 (green vertical block). No recombination is allowed in the translocated fragments that correspond to the first 120 markers of chromosome 1 and markers 350 to 469 of chromosome 4 in the standard structure. The gametic selection deduced in this paper ($p_1 = 1$, $p_2 = 0.07$, $p_3 = 0.1$, $p_4 = 0.4$, $q_1 = 1$, $q_2 = 0$, $q_3 = 0$ and $q_4 = 0.28$) has been applied on this simulated progeny. The linkage between markers is color-coded from strong linkages (warm colors) to no linkages (cold colors). The graph on the upper diagonal of the dot plot represent for each marker the number of heterozygous, homozygous1 and homozygous2 individuals respectively in green, blue, red. The black dot graph represents the marker segregation distortions along the eleven chromosomes. The value presented in the graph is the $-\log_{10}$(p-value of the chi-square test testing the deviation from expected Mendelian segregation ratio). (**A**) Dot-plot with markers organized to the haplotype structure where the red translocated fragment is at the extremity of chromosome 1 and the purple translocated fragment is in chromosome 4. Breakpoints (red arrows) in marker linkage can be observed in the junction between markers located on translocated fragments and markers on flanking region (**B**) Dot-plot with markers organized to the haplotype structure where the translocated fragment is in chromosome 4. No breakpoint is observed.

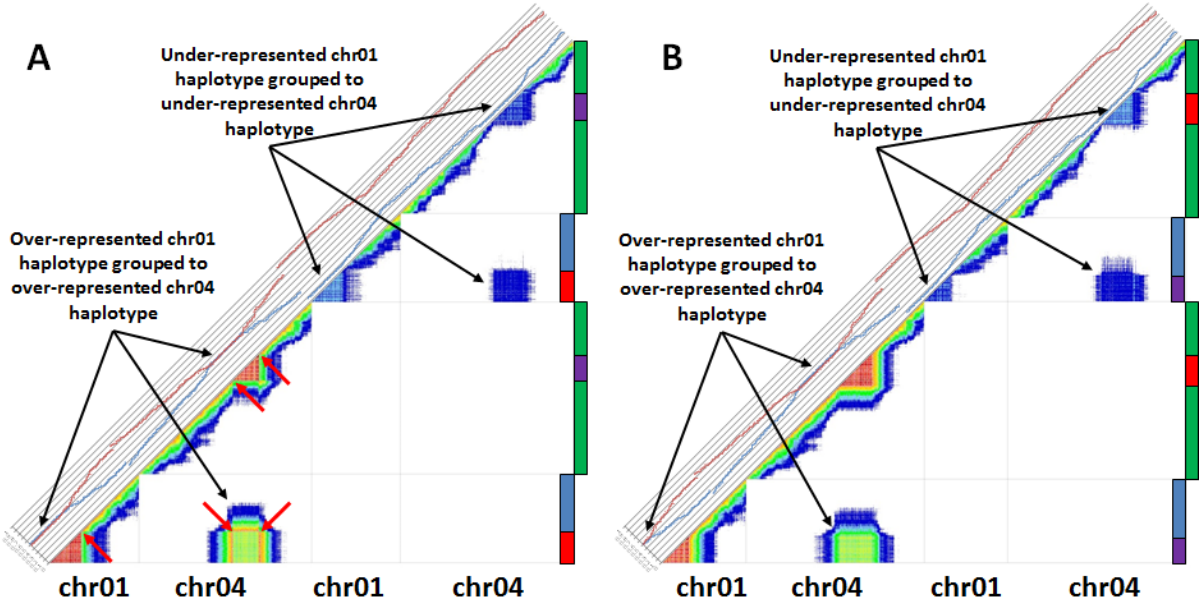


Figure S22: Marker linkage with dominant markers in case of reciprocal translocation with gametic selection model. Dot-plot representing linkages between dominant markers of chromosomes 1 and 4 simulated on a self-progeny of 180 individuals. The simulated chromosomes of the parent of the self-progeny have a reciprocal translocation (red and purple vertical blocks) of the beginning of chromosome 1 (blue vertical block) with a fragment of chromosome 4 (green vertical block). No recombination is allowed in the translocated fragments that correspond to the first 120 markers of chromosome 1 and markers 350 to 469 of chromosome 4 in the standard structure. The gametic selection deduced in this paper ($p_1 = 1$, $p_2 = 0.07$, $p_3 = 0.1$, $p_4 = 0.4$, $q_1 = 1$, $q_2 = 0$, $q_3 = 0$ and $q_4 = 0.28$) has been applied on this simulated progeny. The linkage between markers is color coded from strong linkages (warm colors) to no linkages (cold colors). The graph on the upper diagonal of the dot-plot represents for each marker the proportion of recessive and dominant individuals respectively in blue and red. (**A**) Dot-plot with markers organized to the haplotype structure where the red translocated fragment is at the extremity of chromosome 1 and the purple translocated fragment is in chromosome 4. Breakpoints (red arrows) in marker linkage can be observed in the junction between markers located on translocated fragments and markers on flanking regions (**B**) Dot-plot with markers organized to the haplotype structure where the translocated fragment is in chromosome 4. No breakpoint is observed.

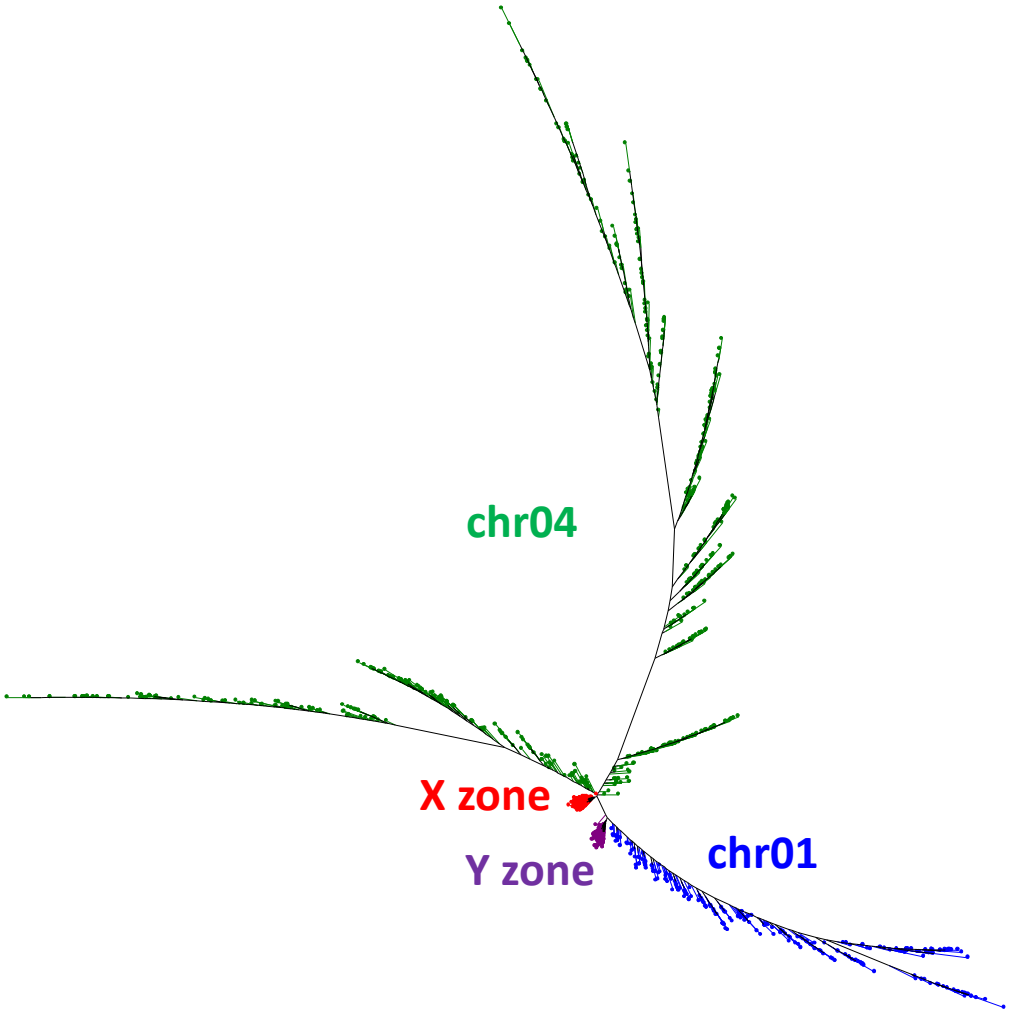


Figure S23: Weighted Neighbor-joining in case of reciprocal translocation with gametic selection model. The tree has been constructed using the pairwise map distances calculated with the Kosambi mapping function on co-dominant markers of chromosomes 1 and 4, simulated on a self-progeny of 180 individuals. The simulated chromosomes of the parent of the self-progeny have a reciprocal translocation of the beginning of chromosome 1 with a fragment of chromosome 4. No recombination is allowed in the translocated fragments that correspond to the first 120 markers of chromosome 1 and markers 350 to 469 of chromosome 4 in the standard structure. The gametic selection deduced in this paper ($p_1 = 1$, $p_2 = 0.07$, $p_3 = 0.1$, $p_4 = 0.4$, $q_1 = 1$, $q_2 = 0$, $q_3 = 0$ and $q_4 = 0.28$) has been applied on this simulated progeny. The distances were calculated from the pairwise recombination frequencies calculated by JoinMap4.1. Markers belonging to chromosome 1 and chromosome 4 are drawn in blue and green respectively. Markers belonging to the translocated fragments are drawn in red and purple respectively.

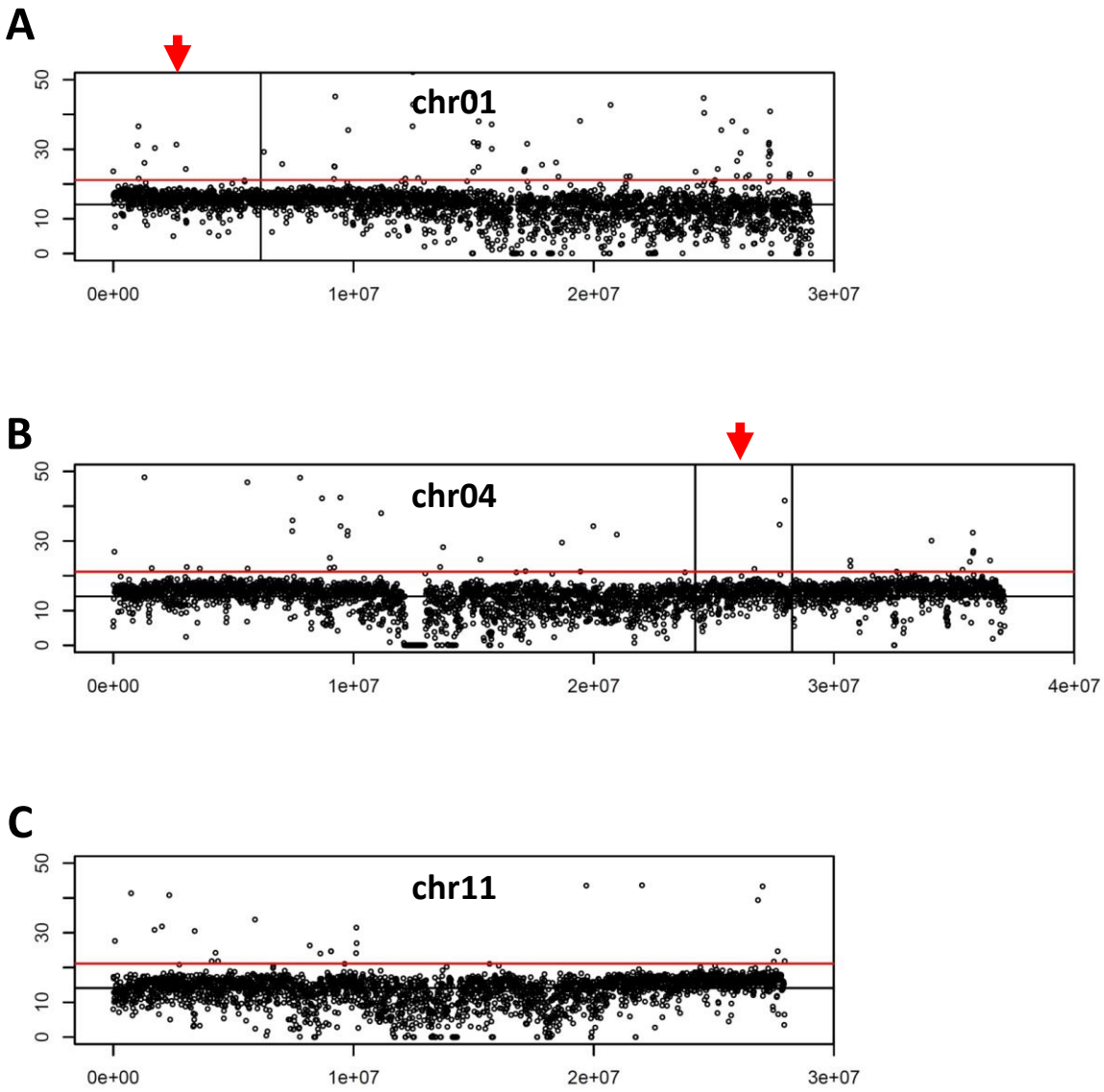


Figure S24: Plot of coverage of 'Pahang' mapped reads along chromosome 1 (**A**), 4 (**B**) and 11 (**C**) of *Musa acuminata* reference sequence. Each dot represents the mean coverage on contiguous windows of 10 kb. The dark lines represent the median coverage calculated from all chromosomes and the red lines represent the expected coverage in case of heterozygous duplication. Red arrows indicate distorted regions where coverage increase is expected in case of duplication.

Conclusion à l'article

L'analyse génétique des marqueurs moléculaires générés sur la population 'Pahang' a révélé deux régions chromosomiques très fortement distordues sur le chromosome 1 et le chromosome 4. Ces régions très distordues co-ségrégent et présentent un très faible taux de recombinaison.

Quatre modèles différents impliquant soit une hétérozygotie de structure (duplication, translocation réciproque ou translocation non réciproque avec possibilité de sélection d'un ou plusieurs gamètes), soit une sélection gamétique seule ont été envisagés pour expliquer ces observations.

Un des résultats principaux est que les hétérozygoties structurales testées ne peuvent pas expliquer, à elles seules, les distorsions de ségrégation observées. La quasi complète létalité de deux des gamètes peut être expliquée par l'absence ou la présence en double copie de certains segments chromosomiques et donc être la conséquence directe d'une hétérozygotie structurale. Par contre, la différence de survie entre les deux gamètes restants nécessite l'introduction d'une sélection génique additionnelle. En introduisant cette sélection gamétique additionnelle, chacun des modèles peut expliquer les données observées.

Sur la base de l'analyse des couvertures du génome de 'Pahang' sur le génome de référence, le modèle de duplication peut être écarté puisqu'aucune hausse significative de couverture n'est observée dans les régions distordues.

Finalement trois modèles, deux impliquant une translocation réciproque ou non avec survie/fitness différentielle des gamètes, et un impliquant quatre gènes ou facteurs génétiques sous sélection, pourraient expliquer les distorsions observées. Le modèle impliquant quatre gènes ou facteurs génétiques n'explique pas, dans sa forme actuelle, les multivalents observés lors de la méiose de 'Pahang'. Ces multivalents n'impliquent néanmoins pas forcément les chromosomes 1 et 4.

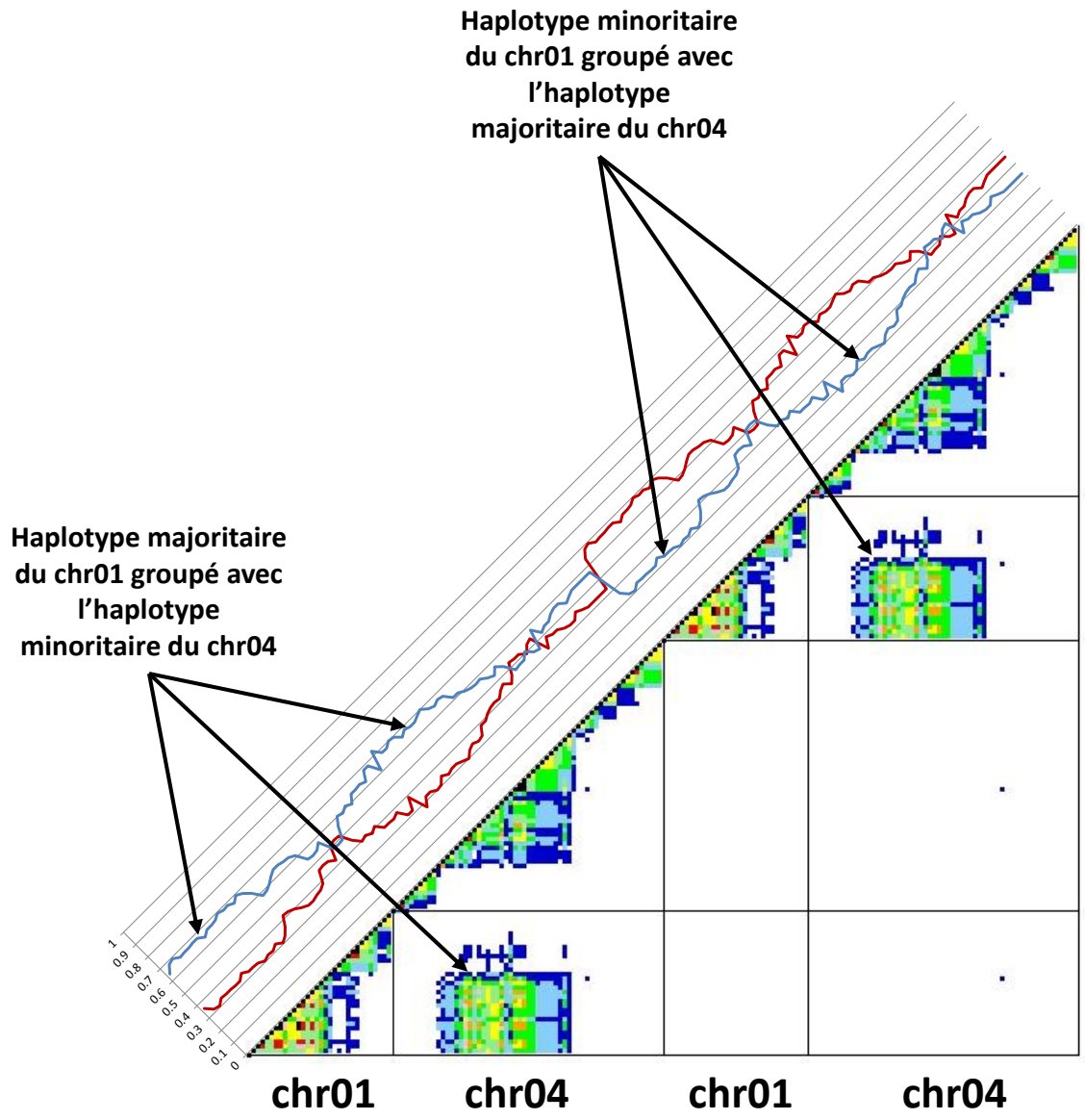


Figure 31 : Proportions alléliques et liaison entre les marqueurs de 'Pisang lilin' observées dans la population Bornéo x Pisang Lilin. Dot-plot présentant la liaison entre les allèles des différents haplotypes des chromosomes 1 et 4. Les fréquences alléliques dont la liaison est observée sont représentées par la courbe bleue. La courbe rouge représente les fréquences alléliques de l'allèle alternatif (ce qui permet de différentier l'allèle majoritaire de l'allèle minoritaire).

Données complémentaires:

L'homologie entre les distorsions observées chez les cartes génétiques 'Pisang lilin' (Hippolyte et al., 2010) et 'Pahang' (D'Hont et al., 2012) suggère que la structure des chromosomes 1 et 4 de 'Pahang' pourrait être la même que 'Pisang lilin'. L'analyse des données de re-séquençage et de génotypage nous permet d'éliminer dans le cas de 'Pahang', le modèle de la duplication proposé pour 'Pisang lilin', c'est à dire une duplication d'un segment du chromosome 1 dans l'une des deux versions du chromosome 4 avec mort du gamète portant le fragment dupliqué à l'état hétérozygote.

Nous avons appliqué le programme de simulation utilisée pour 'Pahang' pour revisiter les données de cartographie obtenues chez 'Pisang lilin'. La représentation simultanée des fréquences alléliques et corrélations entre marqueurs dans la descendance 'Pisang lilin' (**Figure 31**) révèle une liaison entre haplotypes majoritaires des chromosomes 1 et 4 et une liaison entre les haplotypes minoritaire comme dans le cas de 'Pahang'. Ces résultats sont en contradiction avec le modèle de duplication proposé par Hippolyte et al., (2010) (**Figure 32**) qui lui implique la liaison de l'haplotype minoritaire du chromosome 1 avec l'haplotype majoritaire du chromosome 4, la liaison de l'haplotype majoritaire du chromosome 1 avec l'haplotype minoritaire du chromosome 4 et la liaison de l'haplotype majoritaire du chromosome 1 avec l'haplotype majoritaire du chromosome 4. Le modèle qui avait été proposé bien qu'il explique les fréquences alléliques retrouvées dans la descendance de 'Pisang lilin', n'explique pas les fréquences des combinaisons alléliques entre les chromosomes 1 et 4. La différence entre les populations 'Pahang' (autofécondation qui donne accès aux fréquences génotypiques) et 'Pisang lilin' (F1 qui donne accès aux fréquences gamétiques) ne permet pas de transférer directement les paramètres des modèles de 'Pahang' à 'Pisang lilin' mais une approche, similaire à celle réalisée dans cette thèse, de recherche des paramètres de survie de gamètes en fonction de différentes structure supposées de 'Pisang lilin' devrait permettre de conclure si 'Pahang' et 'Pisang lilin' ont une structure chromosomique similaire.

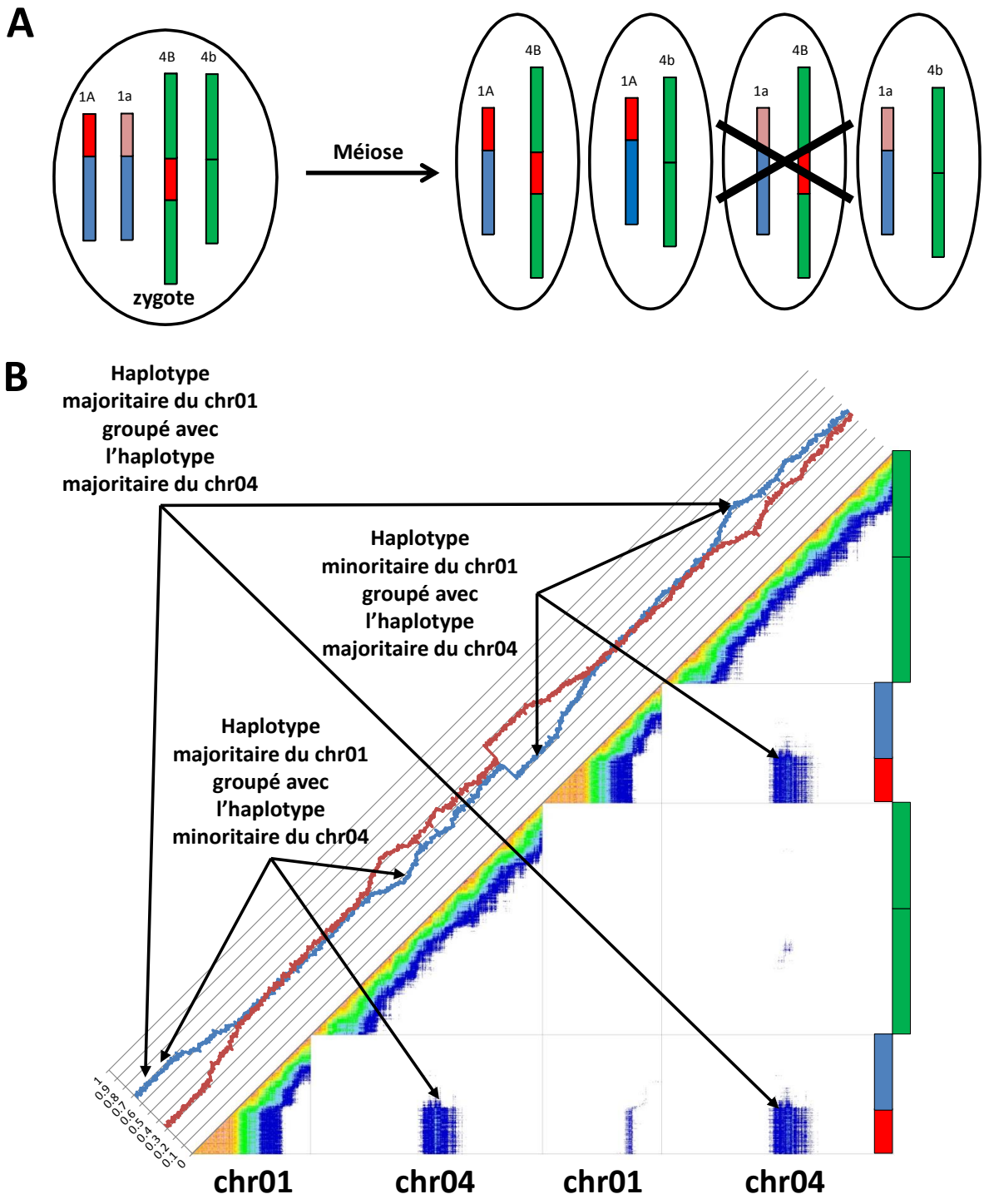


Figure 32 : Simulation du modèle de duplication avec sélection gamétique proposé par Hippolyte et al (2010) pour expliquer les distorsions et pseudo-liaisons observées dans la population Bornéo x Pisang Lilin. (**A**) Modèle de sélection gamétique proposé par Hippolyte et al 2010. (**B**) Dot-plot présentant la liaison entre les allèles des différents haplotypes des deux chromosomes impliqués dans la duplication sur une population simulée selon le modèle proposé en (**A**). Les fréquences alléliques dont la liaison est observée sont représentées par la courbe bleue. La courbe rouge représente les fréquences alléliques de l'allèle alternatif (ce qui permet de différentier l'allèle majoritaire de l'allèle minoritaire).

Conclusion

Lors de ce travail, nous avons mis en place une méthodologie et adapté des outils pour étudier les distorsions de ségrégations qui touchent le génotype ‘Pahang’. Notre approche consiste à observer les liaisons et distorsions d’un grand nombre de marqueurs ordonnés sur la base d’une carte génétique ou d’une séquence de référence, à rechercher des modèles pouvant expliquer les distorsions observées et enfin à valider ces modèles par simulation et comparaisons avec les données réelles. Bien que ce travail n’ait pas permis de trancher de manière précise sur l’origine des distorsions de ségrégations observées chez ‘Pahang’, il a permis une avancée importante dans la caractérisation des distorsions chez ‘Pahang’: dont l’identification précise des zones distordues, la mise en relation avec des zones de recombinaison très réduites et l’identification de paramètres uniques de sélection gamétique selon le modèle envisagé. Ces caractéristiques ont potentiellement des répercussions importantes sur la transmission des caractères (possibilité de combinaison d’alléles ou de recombinaison entre locus) dont les déterminismes génétiques seraient portés par des gènes dans ces zones.

La méthodologie mise en place dans ce travail pourra être poursuivie pour re-visiter le cas de la population issue de ‘Pisang lilin’ mais également pourra être appliquée pour étudier d’autres accessions de bananier voir d’autres espèces présentant des distorsions de ségrégation importantes dans leur descendances.

Discussion générale et perspectives

Ce travail de thèse s'inscrit dans la suite directe du séquençage du génome de bananier qui associé aux nouvelles capacités de séquençage apportées par les NGS, permet d'envisager de nouvelles approches pour différents domaines d'étude chez le bananier. Dans ce contexte, l'objectif fixé au début de ce travail de thèse était de développer des approches basées sur le re-séquençage pour caractériser les variations de structures chromosomiques au sein de l'espèce *Musa acuminata*. Les nouveaux types de données générées par les technologies NGS, de par leur volume notamment, constituent un nouveau challenge et nécessitent le développement de nouvelles approches et de nouveaux outils bioinformatiques et biomathématiques. Ainsi, une part importante du travail de thèse a consisté à concevoir de nouvelles approches et à développer de nouveaux outils bioinformatiques. L'accession 'Pahang' a servi de support pour élaborer, évaluer et adapter les approches de détection de grandes variations structurales et pour l'analyse des ségrégations chromosomiques. Nous avons recherché dans cette accession, avec les méthodes développées, la présence d'hétérozygoties structurales et nous avons tenté d'interpréter les fortes distorsions de ségrégation observées dans sa descendance. L'amélioration de la séquence de référence du bananier s'est avérée nécessaire pour entreprendre la recherche de ces variations structurales à partir de re-séquençage d'accessions et de comparaison avec la séquence de référence. Ce travail a été entrepris avec l'accession 'Pahang-Haploïde Doublé' ('Pahang HD') qui a servi de support pour élaborer, évaluer et adapter les approches d'amélioration des assemblages et d'ancrage de génome de référence.

Dans ce chapitre, nous discuterons les approches et outils mis en place et les résultats obtenus concernant i) l'amélioration de la séquence de référence du bananier ; ii) les ségrégations chromosomiques dans l'accession 'Pahang' en lien avec la structure de ce génome et iii) la recherche de variations structurales par re-séquençage chez le bananier.

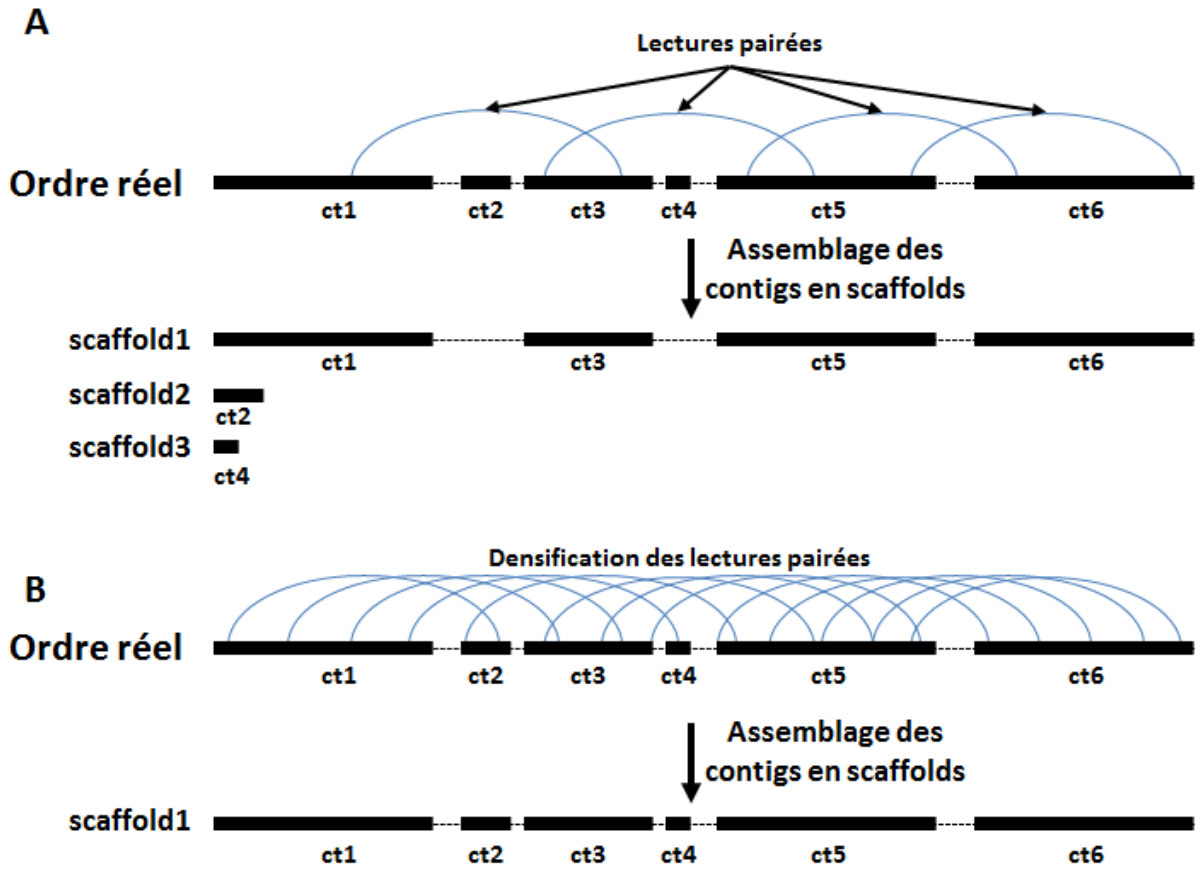


Figure 33 : Assemblage des contigs en scaffolds avec une banque à faible couverture. (**A**) Les lectures pairées ne permettent pas de grouper les petits contigs ct2 et ct4 dans le scaffold1.Cet assemblage comporte un nombre important de scaffold, dont certains de petite taille correspondent aux petits contigs (**B**) La densification des lectures pairées permet de mieux ordonner les contigs et de tous les intégrer dans un seul scaffold.

1. Amélioration de la séquence de référence

Les premiers tests de recherche de variations structurales réalisés à partir de l'alignement de lectures pairées de 'Pahang' sur la séquence de référence ont révélé un très grand nombre de lectures discordantes dues notamment au nombre assez important de petits scaffolds non ancrés. Nous avons donc retravaillé la séquence de référence avec deux objectifs : l'amélioration de l'assemblage et l'amélioration de l'ancrage sur les chromosomes.

Amélioration de l'assemblage

La séquence de référence du bananier (D'Hont et al., 2012) a été construite par une approche de 'whole genome sequencing'. Les contigs, constitués par l'assemblage de lectures 454, ont été assemblés en scaffolds en utilisant des lectures pairées issues de BAC-ends et d'une banque de 10kb. A l'intérieur des scaffolds les contigs sont séparés par des N. Le nombre de N insérés est déterminé sur la base de la taille attendue entre les séquences pairées qui ont servi à assembler les contigs en scaffolds. Cependant ces deux banques représentaient une couverture totale du génome relativement faible (3x au total). Aussi au cours du premier assemblage, un grand nombre de petits scaffolds (inférieurs à 10 kb) et de scaffolds de taille intermédiaires (inférieur à la taille d'un BAC soit environ 100kb) n'avaient pas pu être insérés à leur place notamment à l'intérieur de grands scaffolds (**Figure 33**). Le résultat est la présence de grands scaffolds contenant des régions importantes constitués de N et d'un grand nombre de petits scaffolds (correspondant à ces régions de N) qui ensuite n'ont pu être ancrés aux chromosomes.

Il est intéressant pour l'étape de construction des scaffolds d'avoir à disposition des banques de grande taille. En effet les séquences répétées qui représentent une part importante des génomes (Kejnovsky et al., 2012) sont généralement incorrectement assemblées avec de nombreuses séquences assemblées en un seul contig (Alkan et al., 2011b). Si la taille des banques dont les extrémités sont séquencées n'est pas suffisante pour passer au-dessus des contigs correspondant aux zones répétées (**Figure 34**), l'étape de scaffolding peut poser problème et conduire à des erreurs ou à une fragmentation plus importante de l'assemblage (Pop, 2009).

En vue d'améliorer l'assemblage du génome de bananier de référence, nous avons donc choisi de produire des banques de 5kb à partir de l'accession 'Pahang HD' en utilisant la technologie Illumina. Le choix d'une banque de 5kb peut être discutable puisque la taille moyenne des rétro-éléments à LTR, qui constituent la majorité des séquences répétées du

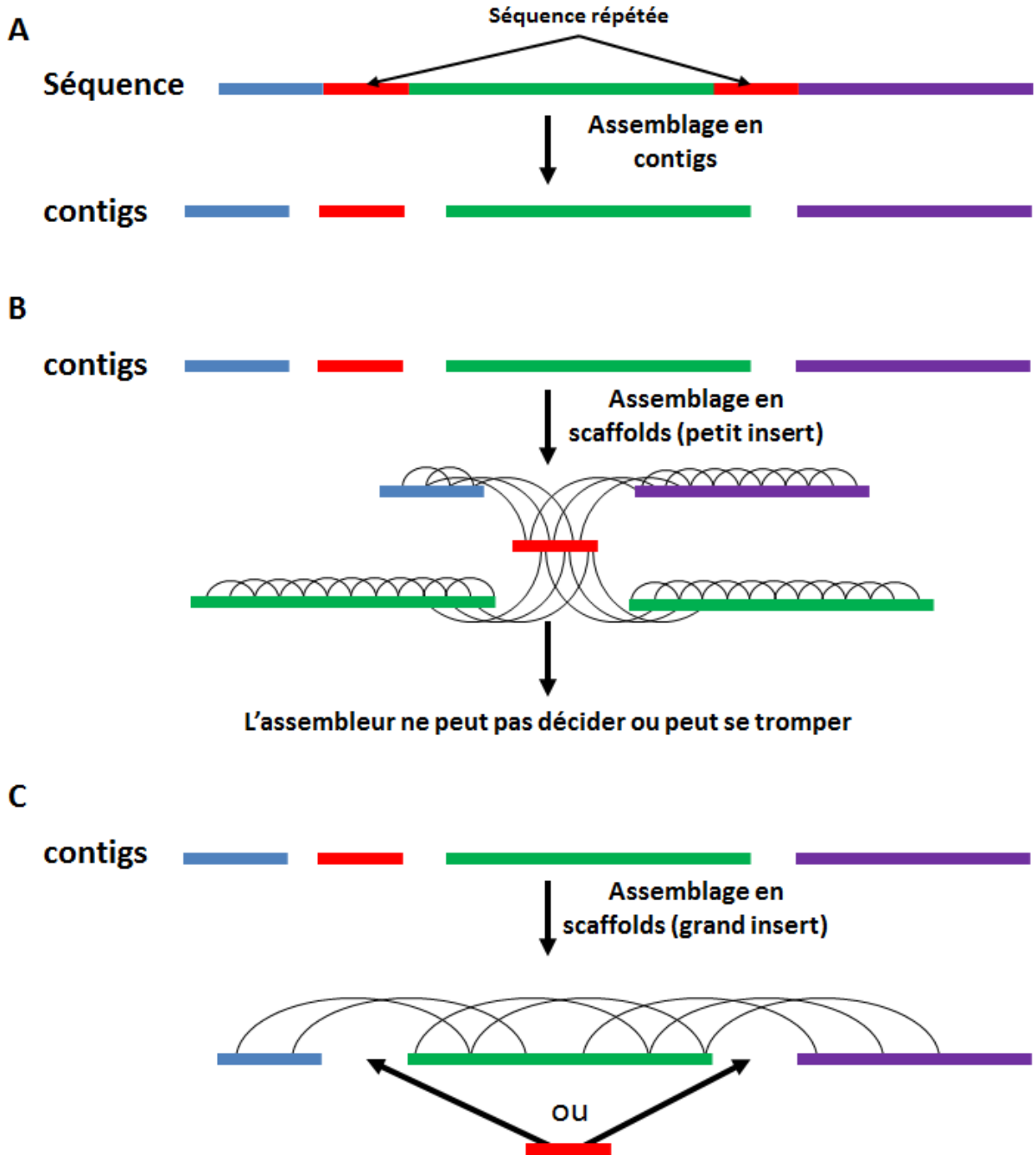


Figure 34 : Assemblage d'une séquence comprenant une région répétée en fonction de la taille de l'insert de la librairie utilisée pour réaliser le séquençage. (**A**) La présence d'une séquence dupliquée (rouge) entraîne des problèmes d'assemblage de la zone séquencée et 4 contigs sont obtenus. (**B**) Les lectures pairées avec insert de petite taille ne permettent pas de passer au-dessus de la zone dupliquée (rouge) et les 4 contigs ne peuvent pas être groupées. (**C**) Les lectures pairées avec insert de grande taille permettent de passer au-dessus de la zone dupliquée (rouge) et donc de grouper les contigs.

génome de la banane (D'Hont et al., 2012), a été estimée à environ 6kb. Dans ce contexte, des banques d'inserts de 8 ou 20 kb auraient semblé plus appropriées. Cependant, plus on augmente la taille des fragments d'ADN des banques, plus les données sont redondantes (Jean-Marc Aury, com. pers.) et donc moins le génome est couvert. Nous avons pu vérifier ceci avec des données générées au sein de l'équipe où (à nombre de lectures égal) une banque de 5kb présente moins de 10% de lectures en plusieurs copies (redondantes), une banque de 8kb présente aux alentours de 40 % de redondance et une banque de 20kb présente 70% de séquences redondantes. Le choix d'une banque de 5kb nous a semblé représenter un bon compromis entre une couverture élevée et une taille "d'insert" élevée.

L'ajout de ce jeu de données supplémentaires, nous a permis en utilisant le programme SSPACE (Boetzer et al., 2011), de générer un nouvel assemblage dans lequel le nombre de scaffolds a été divisé par plus de trois par rapport à la première version de l'assemblage. Il faut cependant noter que cette étape n'a pas entraîné une hausse importante de la N50 des scaffolds (1.3Mb à 1.5Mb) puisqu'une part importante de l'amélioration a consisté en l'inclusion de petits scaffolds à l'intérieur de grands scaffolds.

Pour encore améliorer l'assemblage, nous avons développé des outils qui sont décrits et discutés dans la publication n°1 du chapitre 1. Leurs principaux avantages sont qu'ils sont modulables, qu'ils peuvent s'adapter à différents types de jeux de données. De plus, ils sont semi-automatiques, permettant de faire des compromis locaux par validation manuelle, alors que la plupart des programmes d'assemblage sont basés sur des compromis globaux (à l'échelle du génome) qui ne reflètent pas toujours des réalités locales.

L'amélioration supplémentaire réalisée avec ces outils n'est pas négligeable puisqu'elle a permis de réduire de 20% le nombre de scaffolds mais également de doubler la N50 des scaffolds. L'augmentation de la N50, importante à cette étape, est due au fait que plusieurs scaffolds de grande taille présentaient des problèmes d'assemblage à leurs extrémités qui empêchaient leur jonction avec d'autres scaffolds de grande taille. Ces problèmes d'assemblage ont été identifiés en utilisant les données de cartographie génétique couplées à l'inspection des lectures pairées s'alignant dans les zones problématiques. La correction de ces scaffolds suivie de l'étape de jonction de scaffold a permis de regrouper plusieurs scaffolds de grande taille faisant ainsi augmenter la N50.

A la fin de cette thèse, nous avons également pu disposer d'une carte physique de 'Pahang HD' obtenue par la technologie Irys (http://www.bionanogenomics.com/technology/). Cette technologie est une variante de la technique "d'optical mapping" mais en haut débit. Elle

génère des cartes de restriction sur des fragments d'ADN de haut poids moléculaire. L'arrivée tardive de ces données pendant cette thèse n'a pas permis de les exploiter complètement. Elles n'ont été utilisées que pour réaliser une étape supplémentaire d'assemblage de certains scaffolds avant leur ancrage sur la carte génétique. Cependant, si elles avaient été disponibles, elles auraient pu être utilisées plus tôt dans le processus en particulier pour valider les étapes d'édition de l'assemblage et identifier les problèmes d'assemblage. Plus globalement, ce type de données semble avoir un potentiel important en complément d'autres technologies pour réaliser des assemblages puisque les zones de rupture d'assemblage ne sont pas les mêmes en fonction de la technologie utilisée. Par exemple, on peut penser que les séquences répétées sont moins source de rupture d'assemblage avec la technologie Irys qu'avec les technologies de séquençage.

Amélioration de l'ancrage

Pour augmenter la proportion de l'assemblage ancré aux 11 chromosomes du bananier, nous avons densifié la carte génétique déjà existante de 'Pahang' en y ajoutant plus de 20000 marqueurs obtenus par des approches de génotypage par séquençage (DArTseq). Cependant, ce nombre très élevé de marqueurs, difficilement manipulable par les programmes de cartographie génétique, et couplé aux erreurs de génotypages retrouvées dans ce type de jeu de données (Spindel et al., 2013), nous a conduits à développer notre propre méthodologie pour ancrer les scaffolds. Cette méthodologie a pour principe d'ordonner des groupes de marqueurs entre eux. Un groupe de marqueur est défini par leur appartenance à un scaffold et l'ordre de ces marqueurs dans un groupe est celui de leurs positions sur le scaffold. Un score est calculé sur la base de l'ordre des marqueurs dans les scaffolds ordonnés les uns par rapport aux autres et des tests de permutations de scaffolds permet de trouver un des meilleurs ordres possible. Cette approche permet de travailler avec des jeux de données très importants et également de gérer en partie les erreurs de génotypage. Cependant cette erreur est d'autant mieux gérée que le nombre de marqueurs associés à un scaffold est grand. Ainsi, dans les zones centromériques et péri-centromériques, dont l'assemblage est en général plus fragmenté et la recombinaison moins importante, voire absente (Chen et al., 2002; Gill et al., 1996; Hall et al., 2003; Wu et al., 2003), l'ordre et l'orientation des scaffolds est encore dans notre nouvelle version une approximation du véritable ordre des séquences.

Les outils développés pour améliorer d'une part l'assemblage et d'autre part l'ancrage ont été rendus disponibles et pourront donc être utilisés pour améliorer la séquence d'autres génomes.

Ces programmes et méthodes ont d'ores et déjà été utilisés dans le cadre de la production en cours de la séquence du génome B (*Musa balbisiana*) réalisé en collaboration avec le BGI, pour identifier et corriger des scaffolds et les ordonner en 11 pseudo-molécules. Ces programmes ont également été utilisés pour réaliser l'ancrage de la version CIRAD du génome du palmier à huile (*Elaeis guineensis*) et sont actuellement envisagés pour assembler un génome mosaïque monoploïde de canne à sucre.

Finalement, cet important travail que nous avons réalisé pour améliorer la séquence de référence du génome du bananier, au-delà de la thématique « variation structurale », aura un impact sur l'ensemble des études réalisées par la communauté scientifique qui exploite cette séquence puisque, de la qualité de la séquence de référence dépend la qualité des analyses réalisées à partir de cette référence (Chain et al., 2009).

2-Etude de l'origine des distorsions de ségrégation dans la descendance 'Pahang'

La carte génétique de 'Pahang' réalisée à partir de son autofécondation avait révélé la présence de distorsions de ségrégations importantes impliquant le chromosome 1 et le chromosome 4 (D'Hont et al., 2012) qui pouvait suggérer une hétérozygotie de structure impliquant ces deux chromosomes.

Au cours de cette thèse, la structure chromosomique de l'accession 'Pahang' a donc été étudiée plus en détail en exploitant de nouvelles données de génotypage de cette descendance, des données re-séquençage de l'accession 'Pahang' ainsi qu'en réalisant des simulations.

Quatre hypothèses ou modèles différents dont un n'impliquant que de la sélection gamétique et trois impliquant une hétérozygotie de structure : duplication, translocation réciproque, ou translocation non réciproque chacune avec possibilité de sélection gamétique, ont été envisagés pour expliquer les distorsions de ségrégation et les pseudo-liaisons observées. Pour chacun de ces modèles, nous avons pu trouver des valeurs de sélection gamétique compatibles avec les fréquences génotypiques observées dans la population.

Nous avons montré qu'un modèle d'interaction génique permet d'expliquer les distorsions observées et les sélections gamétiques prédites par le modèle, en l'absence de variations structurales. Ce modèle est théorique et vue la taille importante des zones impliquées dans la distorsion ainsi que la faible recombinaison, il serait difficile d'identifier des gènes ou facteurs génétiques qui pourraient être impliqués dans ce modèle. Par exemple chez le riz, une

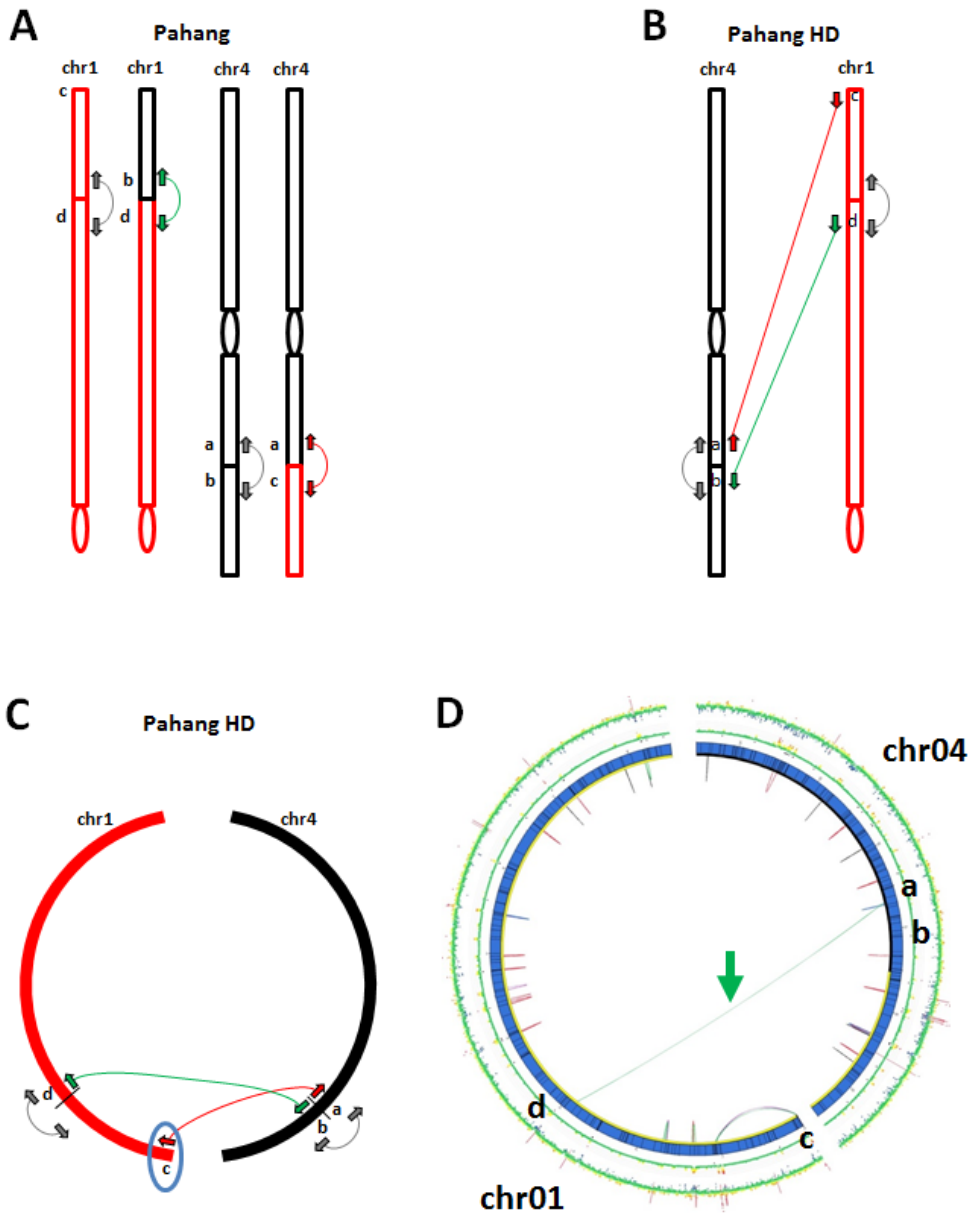


Figure 35 : Hypothèse d'une translocation réciproque. (A) Structure des chromosomes de 'Pahang' représentée avec une translocation réciproque entre les extrémités des chromosomes 1 et 4. La translocation supposée est à l'état hétérozygote. Les lectures pairées obtenues aux bornes des points de réarrangements sont représentées par les flèches violettes, vertes et grises. Dans cette hypothèse, la région c qui est à l'extrémité du chromosome 1 pour l'un des haplotypes se retrouve à l'intérieure du chromosome 4 pour l'autre haplotype. (B) Schématisation de l'alignement des lectures pairées sur le génome de référence. Une partie des lectures pairées (flèches grises) s'alignent de façon concordante car elles correspondent à la structure retrouvée dans Pahang HD. Les autres lectures pairées, flèches rouges et vertes, s'alignent de façon discordantes et relient les régions a à c et b à d respectivement. Ces lectures discordantes correspondent à l'autres haplotype. (C) Schématisation de la représentation Circos attendue de la structure de Pahang telle que présentée en (A) alignée sur le génome de référence comme représenté en (B). En cas d'absence de la région c dans l'assemblage du génome de référence (cercle bleu), les lectures rouges ne peuvent pas être observées et seule la paire de zones discordante verte peut être observée. (D) Représentation Circos des paires de zones discordantes identifiées par alignement de la banque 5 kb de 'Pahang' sur les régions les plus distordues des chromosomes 1 et 4 du génome de référence (*Musa acuminata*). La flèche verte indique la paire de zones discordantes qui relie les régions a et d.

population de 10 000 individus a été nécessaire pour identifier deux zones de 11 et 59 kb sur les chromosomes 6 et 1 respectivement, qui sont responsables d'une incompatibilité génétique entre les sous espèces *indica* et *japonica* chez *Oriza sativa* (Mizuta et al., 2010).

Concernant le modèle impliquant une duplication, l'analyse des données de re-séquençage de l'accession 'Pahang' via une inspection des couvertures obtenues le long des chromosomes 1 et 4 n'a pas montré de segment doublement couvert. Cette absence de hausse de couverture permet d'exclure ce modèle.

Pour les modèles de translocation réciproque et non réciproque, notre pipeline de détection de variations structurales n'a détecté qu'une seule paire de zones discordantes ce qui ne correspond pas à une signature complète d'une translocation (réciproque ou non). Nous avons néanmoins, en fin de rédaction de la thèse, réexaminé en détail ce résultat. Ce qui a conduit à la réflexion qui suit : la localisation de la paire de zones discordantes détectée qui relie les régions distordues des chromosomes 1 et 4 permet, s'il y a effectivement translocation, de supposer que celle-ci implique la partie distale du chromosome 1 et permet également de restreindre les configurations possibles des translocations. Parmi ces configurations, nous avons choisi de nous intéresser au modèle de translocation qui implique une translocation réciproque des extrémités des chromosomes 1 et 4 en orientation inversée (**Figure 35**). En effet ce modèle semble pouvoir expliquer l'absence de recombinaison dans le bras du chromosome 1 et la recombinaison observée sur le bras du chromosome 4 (**Figure 36**). Ainsi, l'orientation inversée par rapport au chromosome 1 du fragment du chromosome 1 transloqué dans le chromosome 4 entraîne à l'issue de la méiose, lorsqu'il y a recombinaison dans ce fragment, la formation d'un chromosome à deux centromères et d'un fragment chromosomique sans centromère correspondant à la région transloquée (**Figure 36 A**). Il est probable que les gamètes portant ces chromosomes ne soient pas viables. Cette hypothèse pourrait donc expliquer l'absence de recombinaisons dans cette région du chromosome 1. Par contre l'orientation identique du fragment transloqué dans le chromosome 4 n'entraîne pas, lorsqu'il y a recombinaison dans cette région, de chromosomes déséquilibrés (**Figure 36 B**). Cette hypothèse pourrait donc expliquer également les recombinaisons, nombreuses, observées dans la région supposée transloquée du chromosome 4 quand on s'éloigne du point de réarrangement.

Ce modèle nécessite deux paires de zones discordantes, dont une paire que nous n'avons pu détecter, reliant l'extrémité du chromosome 1 avec une région contiguë de la zone discordante identifiée dans le chromosome 4. Plusieurs raisons liées à la nature des bornes de la région réarrangée ou à la qualité de l'assemblage de ces bornes développées plus loin dans cette

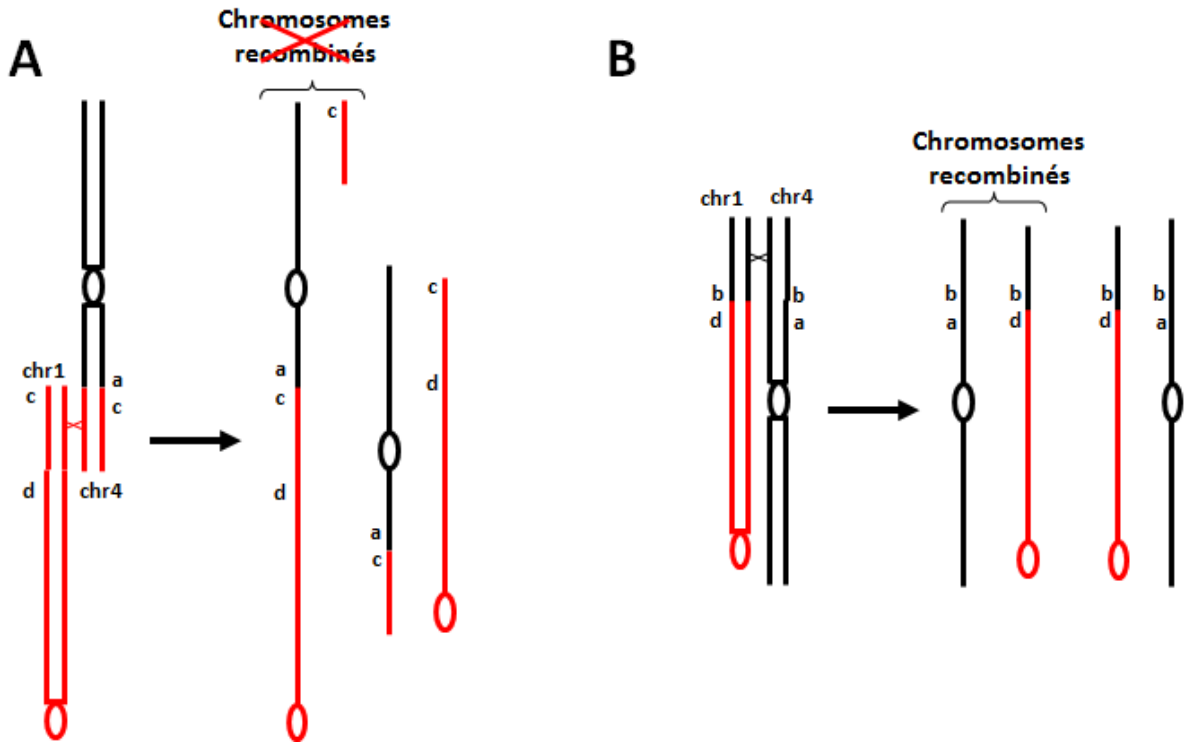


Figure 36 : Schématisation des chromosomes obtenus dans les gamètes en cas de recombinaison entre les fragments transloqués (Cas d'une translocation réciproque entre les extrémités des chromosomes 1 et 4). (A) En cas de recombinaison dans le fragment contenant la région transloquée c, deux types de chromosomes recombinés sont obtenus. Ceux-ci sont fortement déséquilibrés avec un chromosome ayant deux centromères et un chromosome sans aucun centromère. Il est fortement envisageable que ces deux types de chromosomes ne soient pas transmis dans la descendance. (**B**) En cas de recombinaison dans le fragment contenant la région b, deux chromosomes recombinés sont obtenus. Ces deux chromosomes ne sont pas déséquilibrés et devrait donc se retrouver dans la descendance.

discussion (point 3), peuvent être invoquées pour expliquer l'absence de détection de l'ensemble des lectures pairées indicatrices d'une variation structurale. Dans le cas présent, il est probable que notre assemblage ne comprenne pas l'extrémité du chromosome 1 puisque nous n'y retrouvons pas trace de séquences télomériques. L'absence de l'extrémité de ce chromosome pourrait empêcher la détection de la seconde paire de zones discordantes. Pour ces raisons, nous ne pouvons exclure la présence d'une translocation à l'état hétérozygote pour expliquer les distorsions observées. En l'absence de la région télomérique de la séquence de référence, l'inspection des lectures pairées s'alignant dans la zone correspondant à la région "a" de la **Figure 35 C et D** pourrait renseigner sur la validité de ce modèle. En effet, sous ce modèle, soit les lectures sœurs des lectures s'alignant dans cette zone "a" ne s'alignent nulle part, soit elles s'alignent sur un autre des télomères.
Dans tous les cas l'inspection de la séquence de ces lectures sœurs par recherche de séquence télomérique devrait permettre d'étayer ce modèle. Un modèle de translocation a également l'avantage d'être cohérent avec l'observation chez 'Pahang' d'univalents, trivalents et tétravalents sur certaines cellules en metaphase et d'aneuploïdie sur certaines cellules issues de l'anaphase.

Finalement, nous avons montré que le modèle proposé pour 'Pisang Lilin', soit une duplication à l'état hétérozygote entre les chromosomes 1 et 4 (Hippolyte et al., 2010), pour expliquer les distorsions au niveau des chromosomes 1et 4 ne permet pas d'expliquer les distorsions similaires observées chez 'Pahang' et n'explique pas non plus complètement les données de ségrégation de 'Pisang Lilin'.

Des marqueurs PCR chevauchant la zone du réarrangement supposé chez 'Pahang' pourront être développés et testés pour valider cette structure de 'Pahang'. Ces marqueurs pourront également être utilisés pour rechercher la signature de ce réarrangement potentiel chez 'Pisang lilin' et d'autres accessions.

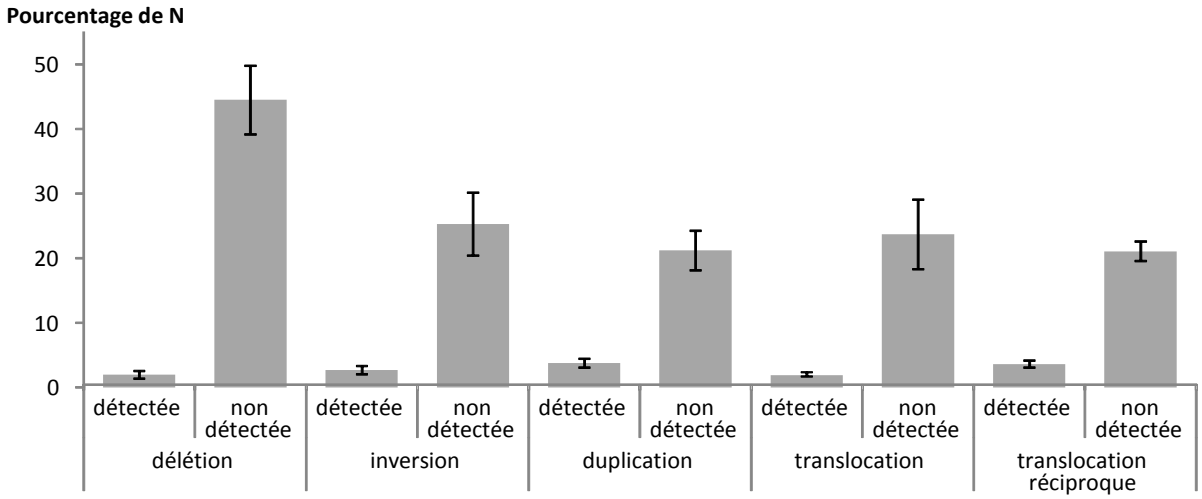


Figure 37 : Détection ou non détection par le pipeline des différents types de variations structurales simulées : Relation avec la proportion (%) de N aux bornes des variations. Pour chaque type de variation structurale, 100 simulations de détection par notre pipeline ont été réalisées à partir d'une banque de couverture 5x. Pour chaque type de variation structurale, la proportion en N sur une fenêtre de plus ou moins 5 kb autour des bornes de la variation structurale a été calculée. La moyenne de cette proportion est représentée en cas de détection ou de non-détection par le pipeline. Les barres d'erreurs correspondent à l'erreur standard.

3-Les approches de re-séquençage pour l'étude des variations structurales chez le bananier : bilan et perspectives

Au cours de cette thèse nous avons développé des approches et outils pour identifier et caractériser les variations structurales pouvant exister entre deux génomes et en particulier entre ceux des sous espèces de *Musa acuminata*. Les premiers tests réalisés avec ces outils ont montré qu'il était essentiel d'améliorer la séquence de référence pour poursuivre ce type d'approche. Ce n'est donc qu'en fin de thèse, une fois la séquence de référence améliorée que nous avons pu re-tester notre approche et nos outils pour rechercher des variations structurales. Nous avons travaillé sur deux accessions, l'accession 'Pahang' chez laquelle on suspecte une variation structurale et l'accession 'PKW' de l'espèce *Musa balbisiana* chez qui le séquençage en cours a montré qu'il existe deux grandes variations par rapport à la référence *Musa acuminata* : une inversion de plusieurs Mb dans le chromosome 5 et une translocation réciproque des extrémités des chromosomes 1 et 3.

Pour l'accession 'Pahang' nous n'avons pas pu mettre en évidence de signature complète d'une variation structurale par contre chez l'accession 'PKW' nous avons pu détecter une des deux variations structurales : l'inversion. Ce résultat sur 'PKW' montre que nos outils sont bien capables d'identifier des variations structurales mais pas toutes.

Plusieurs causes techniques et biologiques peuvent expliquer qu'on a des difficultés à identifier certaines variations structurales avec ce type d'approche. Parmi les raisons techniques, il faut tout d'abord noter que ce type d'approche ne permet d'identifier une variation structurale de grande taille qu'aux bornes de celle-ci ; et plus précisément sur une zone faisant la taille de la banque re-séquencée. La nature de la séquence et la qualité de la séquence de référence dans ces régions impactent donc grandement la détection.

Parmi les raisons techniques liées à la qualité de l'assemblage on peut noter :

- la présence de N dans les zones aux bornes de la variation structurale. Des approches de simulation des différents types de variations structurales identifiées par notre programme sur le chromosome 11 de la nouvelle version de l'assemblage de la banane ont révélées que dans 15% des cas la variation structurale simulée n'est pas détectée. Dans 78% des cas où la variation structurale n'est pas détectée une signature partielle de la variation est observée. L'analyse de ces simulations a révélé une corrélation entre l'absence de détection de la variation structurale et la densité en N aux bornes de la variation simulée (**Figure 37**).

- l'assemblage incomplet des zones aux bornes de la variation structurale. Si un scaffold qui contient une des bornes de la variation structurale se trouve dans la partie de l'assemblage

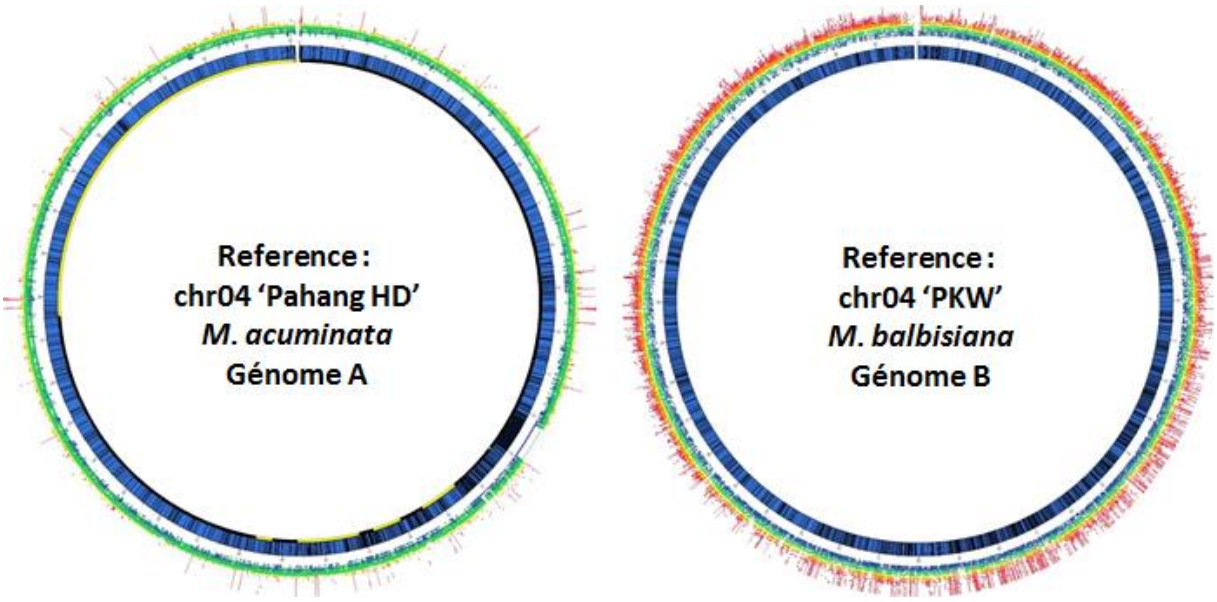


Figure 38 : Distribution des couvertures de la banque 5 kb de 'Pahang' sur les deux génomes de référence *Musa*. Pour chaque de génome de référence de *Musa* (Génomes A et B), la banque 5k générée sur l'accession 'Pahang' (génome A) a été alignée et les couvertures calculées. Les couvertures fluctuent de manière importante sur le génome B alors qu'elles sont beaucoup plus homogènes sur le génome A.

qui n'a pas pu être ancrée sur un chromosome, le lien qui relie cette borne de la variation structurale avec l'autre borne pointera vers la partie non ancrée. Comme la partie non ancrée regroupe les scaffolds sans ordre, il n'est alors pas possible d'identifier la variation structurale.

Parmi les raisons biologiques on peut noter :

- la présence de séquences répétées dans les zones aux bornes de la variation structurale. Les séquences répétées, de parts leur nature mobile dans les génomes (McClintock, 1950) peuvent générer un grand nombre de lectures discordantes quand le génome re-séquencé est aligné sur le génome de référence. De plus, ces séquences répétées sont fréquement retrouvées aux bornes des variations structurales (Carbone et al., 2014; Girirajan et al., 2009) et pourraient, dans certains cas, être impliquées dans ces réarrangements. L'impact de ces mouvements et l'amplification différentiel de ces séquences répétées est variable selon la divergence entre le génome re-séquencé et le génome qui sert de référence. Une accession proche du génome de référence présente des variations en séquences répétées moindre qu'une accession plus divergente (Novák et al., 2014). Cependant, l'évolution rapide de ces séquences répétées doit faire qu'elles sont moins un problème si les génomes comparés sont suffisamment divergents puisque dans ce cas les lectures pairées correspondant à des séquences qui ont beaucoup divergées, ne s'alignent plus sur le génome de référence.

Ainsi dans le cas de l'accession 'PKW', la présence de séquences répétées aux points de réarrangements semble être à l'origine de la non détection par notre pipeline de la signature de la translocation réciproque.

- la présence de régions réarrangées dans les zones aux bornes de la variation structurale. Chez le Gibbon par exemple, il a été montré que les bornes de certaines des variations structurales présentaient une structure mosaïque riche en séquences répétées et réarrangées par rapport au génome qui est utilisé comme référence (Girirajan et al., 2009). Cette structure peut perturber l'alignement des lectures et ainsi empêcher la détection de lectures discordantes.

- une divergence importante entre le génome de référence et le génome séquencé, qui présente l'avantage d'atténuer les problèmes liés aux séquences répétées mais qui entraîne un moins bon et moins homogène alignement de l'ensemble des lectures. Ainsi, la distribution des couvertures est bien moins homogène dans le cas de l'alignement de lectures pairées de l'accession 'Pahang HD' sur la sequence de réference du génome B que dans le cas de leur alignement sur le génome A ('Pahang HD') (**Figure 38**).

- une autre difficulté peut venir de la **nature hétérozygote de structure de l'accession re-séquençée** qui réduit de moitié le nombre de lectures qui permettent de détecter la variation par rapport à un homozygote de structure. Pour pallier à cette éventualité, le re-séquençage de tels accessions nécessite de doubler la couverture de séquençage qui aurait été nécessaire pour une accession homozygote de structure. Ce point est un inconvénient dans le cas de l'étude de l'accession 'Pahang' qui est suspectée hétérozygote de structure. Cependant ce défaut est pallié par sa très forte proximité avec le génome de référence.

Pour les raisons biologiques et des raisons de qualité d'assemblage expliquées plus haut, il sera probablement difficile de pouvoir avec notre outil, au moins directement, identifier toutes les variations structurales. Dans ce contexte, d'autres approches seules ou combinées avec éventuellement une restriction pas à pas de la zone investiguée pourrait être envisagée. Ce type d'approche pourrait comprendre des approches de chromosomes painting qui ont été utilisées pour caractériser les différences de structure chromosomique entre deux espèces d'*Arabidopsis* (Lysak et al., 2006). Ce type d'approche, qui est en cours d'évaluation au sein de l'équipe en collaboration avec l'équipe tchèque de Martin Lysak, si elle est applicable aux espèces *Musa*, pourrait permettre d'identifier la présence de variations structurales de grandes tailles et également d'identifier les chromosomes impliqués dans ces variations. La cartographie haute densité peut aussi être envisagée pour restreindre la zone de recherche de la variation structurale en utilisant les approches décrites dans le chapitre III de cette thèse. Une analyse de cartographie physique en utilisant la technologie Irys peut également être envisagée, au moins pour les variations structurales à l'état homozygote. Enfin, l'approche de re-séquençage peut être envisagée pour analyser finement les lectures discordantes dans les zones cibles et pour tenter de localiser finement les bornes de la variation structurale.
Même si notre approche n'est pas parfaite, elle a montré qu'elle permettait de détecter certaines variations structurales. Il sera donc intéressant de la tester sur un nombre plus important d'accessions appartenant aux différents "groupes de translocations" supposés chez les bananiers (Shepherd, 1999). Des données de re-séquengaçe de banques à grand insert chez des accessions appartenant à ces groupes sont nouvellement disponibles mais nous n'avons pu les traiter dans le cadre de la thèse faute de temps.

Un intérêt important de l'approche par re-séquençage de la recherche de variations structurales par rapport à d'autres approches est théoriquement de permettre d'identifier directement les régions chromosomiques impliquées dans la variation et d'atteindre une

résolution fine des points de réarrangement. Avoir une résolution fine des points de réarrangement permet d'envisager le développement de marqueurs diagnostic de la variation de structure (de type PCR). Ceci pourrait permettre d'étudier plus facilement la structure d'autres accessions et de superposer cette diversité de structure des génomes avec les patrons de diversité génétique (Perrier et al., 2011). Ces variations pourront alors être re-situées dans l'histoire des bananiers: dans leur contexte de spéciation en cours et dans le contexte de l'émergence des cultivars via les hybridations intersubspécifiques. Il est également important de noter que l'espèce *Musa acuminata* semble particulièrement prône aux variations structurales. Il serait intéressant de connaître les mécanismes à l'origine de ces variations. L'étude des séquences aux points de réarrangements identifiés pourra fournir des éléments de réponse.

Par ailleurs, chez le bananier, ces variations de structure semblent avoir un impact important sur les recombinaisons et les ségrégations chromosomiques. Une meilleure caractérisation de ces variations de structures chromosomiques, de leur impact sur la transmissions des chromosomes, et donc des caractères, et leur impact sur la stérilité est essentielle pour orienter les programmes d'amélioration variétale.

Bibliographie

Abyzov, A., Urban, A.E., Snyder, M., and Gerstein, M. (2011). CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Res. 21, 974–984.

Adams, K.L., and Wendel, J.F. (2005). Polyploidy and genome evolution in plants. Curr. Opin. Plant Biol. 8, 135–141.

Albers, C.A., Lunter, G., MacArthur, D.G., McVean, G., Ouwehand, W.H., and Durbin, R. (2011). Dindel: Accurate indel calls from short-read data. Genome Res. 21, 961–973.

Alkan, C., Kidd, J.M., Marques-Bonet, T., Aksay, G., Antonacci, F., Hormozdiari, F., Kitzman, J.O., Baker, C., Malig, M., Mutlu, O., et al. (2009). Personalized copy number and segmental duplication maps using next-generation sequencing. Nat. Genet. 41, 1061–1067.

Alkan, C., Coe, B.P., and Eichler, E.E. (2011a). Genome structural variation discovery and genotyping. Nat. Rev. Genet. 12, 363–376.

Alkan, C., Sajjadian, S., and Eichler, E.E. (2011b). Limitations of next-generation genome sequence assembly. Nat. Methods 8, 61–65.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. 215, 403–410.

Alverson, A.J., Rice, D.W., Dickinson, S., Barry, K., and Palmer, J.D. (2011). Origins and Recombination of the Bacterial-Sized Multichromosomal Mitochondrial Genome of Cucumber. Plant Cell Online 23, 2499–2513.

Bakry, F. (2008). Zygotic embryo rescue in bananas. Fruits 63, 111–115.

Bakry, F., Carreel, F., Jenny, C., and Horry, J.-P. (2009). Genetic Improvement of Banana. In Breeding Plantation Tree Crops: Tropical Species, S.M. Jain, and P.M. Priyadarshan, eds. (Springer New York), pp. 3–50.

Bashir, A., Volik, S., Collins, C., Bafna, V., and Raphael, B.J. (2008). Evaluation of Paired-End Sequencing Strategies for Detection of Genome Rearrangements in Cancer. PLoS Comput Biol 4, e1000051.

Bashir, A., Bansal, V., and Bafna, V. (2010). Designing deep sequencing experiments: detecting structural variation and estimating transcript abundance. BMC Genomics 11, 385.

Bikard, D., Patel, D., Le Metté, C., Giorgi, V., Camilleri, C., Bennett, M.J., and Loudet, O. (2009). Divergent Evolution of Duplicate Genes Leads to Genetic Incompatibilities Within A. thaliana. Science 323, 623–626.

Birchler, J.A., Dawe, R.K., and Doebley, J.F. (2003). Marcus Rhoades, Preferential Segregation and Meiotic Drive. Genetics 164, 835–841.

Boetzer, M., Henkel, C.V., Jansen, H.J., Butler, D., and Pirovano, W. (2011). Scaffolding pre-assembled contigs using SSPACE. Bioinformatics 27, 578–579.

Boeva, V., Zinovyev, A., Bleakley, K., Vert, J.-P., Janoueix-Lerosey, I., Delattre, O., and Barillot, E. (2011). Control-free calling of copy number alterations in deep-sequencing data using GC-content normalization. Bioinformatics 27, 268–269.

Brown, J.D., and O'Neill, R.J. (2009). The Mysteries of Chromosome Evolution in Gibbons: Methylation Is a Prime Suspect. PLoS Genet. 5, e1000501.

Buckler, E.S., Phelps-Durr, T.L., Buckler, C.S.K., Dawe, R.K., Doebley, J.F., and Holtsford, T.P. (1999). Meiotic Drive of Chromosomal Knobs Reshaped the Maize Genome. Genetics 153, 415–426.

Cañestro, C., Albalat, R., Irimia, M., and Garcia-Fernàndez, J. (2013). Impact of gene gains, losses and duplication modes on the origin and diversification of vertebrates. Semin. Cell Dev. Biol. 24, 83–94.

Carbone, L., Vessere, G.M., Hallers, B.F.H. ten, Zhu, B., Osoegawa, K., Mootnick, A., Kofler, A., Wienberg, J., Rogers, J., Humphray, S., et al. (2006). A High-Resolution Map of Synteny Disruptions in Gibbon and Human Genomes. PLoS Genet. 2, e223.

Carbone, L., Harris, R.A., Vessere, G.M., Mootnick, A.R., Humphray, S., Rogers, J., Kim, S.K., Wall, J.D., Martin, D., Jurka, J., et al. (2009). Evolutionary Breakpoints in the Gibbon Suggest Association between Cytosine Methylation and Karyotype Evolution. PLoS Genet. 5, e1000538.

Carbone, L., Alan Harris, R., Gnerre, S., Veeramah, K.R., Lorente-Galdos, B., Huddleston, J., Meyer, T.J., Herrero, J., Roos, C., Aken, B., et al. (2014). Gibbon genome and the fast karyotype evolution of small apes. Nature 513, 195–201.

Carreel, F. (1994). Etude de la diversité génétique des bananiers (genre Musa) à l'aide des marqueurs RFLP. Institut National Agronomique Paris-Grignon.

Carreel, F., Fauré, S., De Leon, D.G., Lagoda, P.J.L., Perrier, X., Bakry, F., Du Montcel, H.T., Lanaud, C., and Horry, J.P. (1994). Evaluation de la diversité génétique chez les bananiers diploïdes (Musa sp). Genet. Sel. Evol. 26, 125s – 136s.

Carreel, F., de Leon, D.G., Lagoda, P., Lanaud, C., Jenny, C., Horry, J.P., and du Montcel, H.T. (2002). Ascertaining maternal and paternal lineage within Musa by chloroplast and mitochondrial DNA RFLP analyses. Genome 45, 679–692.

Carrier, G. (2011). Bases moléculaires de la variation clonale chez la vigne (Vitis vinifera L.): approche pangénomique. Montpellier, SupAgro.

Chain, P.S.G., Grafham, D.V., Fulton, R.S., FitzGerald, M.G., Hostetler, J., Muzny, D., Ali, J., Birren, B., Bruce, D.C., Buhay, C., et al. (2009). Genome Project Standards in a New Era of Sequencing. Science 326, 236–237.

Champion, J. (1963). Le bananier: techniques agricoles et productions tropicales (Maisonneuve & Larose).

Cheesman, E.E. (1947). Classification of the Bananas: The Genus Musa L. Kew Bull. 106–117.

Chen, K., Wallis, J.W., McLellan, M.D., Larson, D.E., Kalicki, J.M., Pohl, C.S., McGrath, S.D., Wendl, M.C., Zhang, Q., Locke, D.P., et al. (2009). BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nat. Methods 6, 677–681.

Chen, M., Presting, G., Barbazuk, W.B., Goicoechea, J.L., Blackmon, B., Fang, G., Kim, H., Frisch, D., Yu, Y., Sun, S., et al. (2002). An Integrated Physical and Genetic Map of the Rice Genome. Plant Cell Online 14, 537–545.

Chiang, D.Y., Getz, G., Jaffe, D.B., O'Kelly, M.J.T., Zhao, X., Carter, S.L., Russ, C., Nusbaum, C., Meyerson, M., and Lander, E.S. (2009). High-resolution mapping of copy-number alterations with massively parallel sequencing. Nat. Methods 6, 99–103.

Christelova, P., Valarik, M., Hribova, E., De Langhe, E., and Dolezel, J. (2011). A multi gene sequence-based phylogeny of the Musaceae (banana) family. BMC Evol. Biol. 11, 103.

Chumley, T.W., Palmer, J.D., Mower, J.P., Fourcade, H.M., Calie, P.J., Boore, J.L., and Jansen, R.K. (2006). The Complete Chloroplast Genome Sequence of Pelargonium × hortorum: Organization and Evolution of the Largest and Most Highly Rearranged Chloroplast Genome of Land Plants. Mol. Biol. Evol. 23, 2175–2190.

Cui, L., Wall, P.K., Leebens-Mack, J.H., Lindsay, B.G., Soltis, D.E., Doyle, J.J., Soltis, P.S., Carlson, J.E., Arumuganathan, K., Barakat, A., et al. (2006). Widespread genome duplications throughout the history of flowering plants. Genome Res. 16, 738–749.

Cusack, B.P., and Wolfe, K.H. (2007). Not Born Equal: Increased Rate Asymmetry in Relocated and Retrotransposed Rodent Gene Duplicates. Mol. Biol. Evol. 24, 679–686.

Derrien, T., Estellé, J., Marco Sola, S., Knowles, D.G., Raineri, E., Guigó, R., and Ribeca, P. (2012). Fast Computation and Applications of Genome Mappability. PLoS ONE 7, e30377.

Dessauw, D. (1987). Etude des facteurs de la stérilité du bananier (Musa spp.) et des relations cytotaxinomiques entre M. acuminata Colla et M. balbisiana Colla. Université de Paris-sud.

D'hont, A. (2005). Unraveling the genome structure of polyploids using FISH and GISH; examples of sugarcane and banana. Cytogenet. Genome Res. 109, 27–33.

D'Hont, A., Quetier, F., Teoule, E., and Dattee, Y. (1987). Mitochondrial and chloroplast DNA analysis of interspecific somatic hybrids of a leguminosae: Medicago (alfalfa). Plant Sci. 53, 237–242.

D'Hont, A., Paget-Goy, A., Escoute, J., and Carreel, F. (2000). The interspecific genome structure of cultivated banana, Musa spp. revealed by genomic DNA in situ hybridization. Theor. Appl. Genet. 100, 177–183.

D'Hont, A., Denoeud, F., Aury, J.-M., Baurens, F.-C., Carreel, F., Garsmeur, O., Noel, B., Bocs, S., Droc, G., Rouard, M., et al. (2012). The banana (Musa acuminata) genome and the evolution of monocotyledonous plants. Nature 488, 213–217.

Dita, M.A., Waalwijk, C., Buddenhagen, I.W., Souza Jr, M.T., and Kema, G.H.J. (2010). A molecular diagnostic for tropical race 4 of the banana fusarium wilt pathogen. Plant Pathol. 59, 348–357.

Dobzhansky, T. (1937). Genetics and the Origin of Species (Columbia University Press).

Dobzhansky, T., and Sturtevant, A.H. (1938). Inversions in the chromosomes of Drosophila pseudoobscura. Genetics 23, 28.

Dodds, K.S. (1943). Genetical and cytological studies of Musa. V. Certain edible diploids. J. Genet. 45, 113–138.

Dodds, K., and Simmonds, N. (1948). Sterility and parthenocarpy in diploid hybrids of musa. Heredity 2, 101–117.

Dohm, J.C., Lottaz, C., Borodina, T., and Himmelbauer, H. (2008). Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. Nucleic Acids Res. 36, e105–e105.

Dubcovsky, J., Luo, M.-C., Zhong, G.-Y., Bransteitter, R., Desai, A., Kilian, A., Kleinhofs, A., and Dvořák, J. (1996). Genetic Map of Diploid Wheat, Triticum monococcum L., and Its Comparison With Maps of Hordeum vulgare L. Genetics 143, 983–999.

Emde, A.-K., Schulz, M.H., Weese, D., Sun, R., Vingron, M., Kalscheuer, V.M., Haas, S.A., and Reinert, K. (2012). Detecting genomic indel variants with exact breakpoints in single- and paired-end sequencing data using SplazerS. Bioinformatics 28, 619–627.

Fang, Y., Wu, H., Zhang, T., Yang, M., Yin, Y., Pan, L., Yu, X., Zhang, X., Hu, S., Al-Mssallem, I.S., et al. (2012). A Complete Sequence and Transcriptomic Analyses of Date Palm (Phoenix dactylifera L.) Mitochondrial Genome. PLoS ONE 7, e37164.

Fauré, S. (1993). Contribution à la cartographie génétique du génome des bananiers diploides à l'aide de marqueurs moléculaires. Ecole nationale supérieure agronomique de rennes.

Fauré, S., Bakry, F., and González de Leon, D. (1993a). Cytogenetic studies of diploid bananas. In Breeding Banana and Plantain for Resistance to Diseases and Pests, (CIRAD-FLHOR, Montpellier: Ganry J.), pp. 77–92.

Fauré, S., Noyer, J.L., Horry, J.P., Bakry, F., Lanaud, C., and Gońzalez de León, D. (1993b). A molecular marker-based linkage map of diploid bananas (Musa acuminata). Theor. Appl. Genet. 87, 517–526.

Fauré, S., Noyer, J.-L., Carreel, F., Horry, J.-P., Bakry, F., and Lanaud, C. (1994). Maternal inheritance of chloroplast genome and paternal inheritance of mitochondrial genome in bananas (Musa acuminata). Curr. Genet. 25, 265–269.

Feuillet, C., Leach, J.E., Rogers, J., Schnable, P.S., and Eversole, K. (2011). Crop genome sequencing: lessons and rationales. Trends Plant Sci. 16, 77–88.

Feulner, P.G.D., Chain, F.J.J., Panchal, M., Eizaguirre, C., Kalbe, M., Lenz, T.L., Mundry, M., Samonte, I.E., Stoll, M., Milinsky, M., et al. (2013). Genome-wide patterns of standing genetic variation in a marine population of three-spined sticklebacks. Mol. Ecol. 22, 635–649.

Fishman, L., Kelly, A.J., Morgan, E., and Willis, J.H. (2001). A Genetic Map in the Mimulus guttatus Species Complex Reveals Transmission Ratio Distortion due to Heterospecific Interactions. Genetics 159, 1701–1716.

Garsmeur, O., Schnable, J.C., Almeida, A., Jourda, C., D'Hont, A., and Freeling, M. (2014). Two Evolutionarily Distinct Classes of Paleopolyploidy. Mol. Biol. Evol. 31, 448–454.

Gill, K.S., Gill, B.S., Endo, T.R., and Taylor, T. (1996). Identification and high-density mapping of gene-rich regions in chromosome group 1 of wheat. Genetics 144, 1883–1891.

Girirajan, S., Chen, L., Graves, T., Marques-Bonet, T., Ventura, M., Fronick, C., Fulton, L., Rocchi, M., Fulton, R.S., Wilson, R.K., et al. (2009). Sequencing human–gibbon breakpoints of synteny reveals mosaic new insertions at rearrangement sites. Genome Res. 19, 178–190.

Gong, Q., Tao, Y., Yang, J.-R., Cai, J., Yuan, Y., Ruan, J., Yang, J., Liu, H., Li, W., Lu, X., et al. (2013). Identification of medium-sized genomic deletions with low coverage, mate-paired restricted tags. BMC Genomics 14, 51.

Goureau, A., Yerle, M., Schmitz, A., Riquet, J., Milan, D., Pinton, P., Frelat, G., and Gellin, J. (1996). Human and Porcine Correspondence of Chromosome Segments Using Bidirectional Chromosome Painting. Genomics 36, 252–262.

Gray, Y.H.M. (2000). It takes two transposons to tango:transposable-element-mediated chromosomal rearrangements. Trends Genet. 16, 461–468.

Gribble, S.M., Fiegler, H., Burford, D.C., Prigmore, E., Yang, F., Carr, P., Ng, B.L., Sun, T., Kamberov, E.S., Makarov, V.L., et al. (2004). Applications of combined DNA microarray and chromosome sorting technologies. Chromosome Res. 12, 35–43.

Gribble, S.M., Ng, B.L., Prigmore, E., Fitzgerald, T., and Carter, N.P. (2009). Array painting: a protocol for the rapid analysis of aberrant chromosomes using DNA microarrays. Nat. Protoc. 4, 1722–1736.

Guisinger, M., Chumley, T., Kuehl, J., Boore, J., and Jansen, R. (2010). Implications of the Plastid Genome Sequence of Typha (Typhaceae, Poales) for Understanding Genome Evolution in Poaceae. J. Mol. Evol. 70, 149–166.

Hall, S.E., Kettler, G., and Preuss, D. (2003). Centromere Satellites From Arabidopsis Populations: Maintenance of Conserved and Variable Domains. Genome Res. 13, 195–205.

Hampton, O.A., Den Hollander, P., Miller, C.A., Delgado, D.A., Li, J., Coarfa, C., Harris, R.A., Richards, S., Scherer, S.E., Muzny, D.M., et al. (2009). A sequence-level map of chromosomal breakpoints in the MCF-7 breast cancer cell line yields insights into the evolution of a cancer genome. Genome Res. 19, 167–177.

Handsaker, R.E., Korn, J.M., Nemesh, J., and McCarroll, S.A. (2011). Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. Nat. Genet. 43, 269–276.

Hart, S.N., Sarangi, V., Moore, R., Baheti, S., Bhavsar, J.D., Couch, F.J., and Kocher, J.-P.A. (2013). SoftSearch: Integration of Multiple Sequence Features to Identify Breakpoints of Structural Variations. PLoS ONE 8, e83356.

Hippolyte, I., Bakry, F., Seguin, M., Gardes, L., Rivallan, R., Risterucci, A.-M., Jenny, C., Perrier, X., Carreel, F., Argout, X., et al. (2010). A saturated SSR/DArT linkage map of Musa acuminata addressing genome rearrangements among bananas. BMC Plant Biol. 10, 65.

Hippolyte, I., Jenny, C., Gardes, L., Bakry, F., Rivallan, R., Pomies, V., Cubry, P., Tomekpe, K., Risterucci, A.M., Roux, N., et al. (2012). Foundation characteristics of edible Musa triploids revealed from allelic distribution of SSR markers. Ann. Bot.

Hormozdiari, F., Alkan, C., Eichler, E.E., and Sahinalp, S.C. (2009). Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. Genome Res. 19, 1270–1278.

Hormozdiari, F., Hajirasouliha, I., McPherson, A., Eichler, E.E., and Sahinalp, S.C. (2011). Simultaneous structural variation discovery among multiple paired-end sequenced genomes. Genome Res. 21, 2203–2212.

Hřibová, E., Neumann, P., Matsumoto, T., Roux, N., Macas, J., and Dolezel, J. (2010). Repetitive part of the banana (Musa acuminata) genome investigated by low-depth 454 sequencing. BMC Plant Biol. 10, 204.

Iovene, M., Zhang, T., Lou, Q., Buell, C.R., and Jiang, J. (2013). Copy number variation in potato – an asexually propagated autotetraploid species. Plant J. 75, 80–89.

Ivakhno, S., Royce, T., Cox, A.J., Evers, D.J., Cheetham, R.K., and Tavare, S. (2010). CNAseg--a novel framework for identification of copy number changes in cancer from second-generation sequencing data. Bioinformatics 26, 3051–3058.

Jackman, S., Raymond, A., Vandervalk, B., Mohamadi, H., Warren, R., Pleasance, S., Coope, R., Yuen, M.M., Keeling, C., Ritland, C., et al. (2014). Assembling the genomes of the plastid and mitochondrion of white spruce (Picea glauca).

Jáuregui, B., de Vicente, M.C., Messeguer, R., Felipe, A., Bonnet, A., Salesses, G., and Arús, P. (2001). A reciprocal translocation between 'Garfi' almond and 'Nemared' peach. Theor. Appl. Genet. 102, 1169–1176.

Jenczewski, E., Gherardi, M., Bonnin, I., Prosperi, J.M., Olivieri, I., and Huguet, T. (1997). Insight on segregation distortions in two intraspecific crosses between annual species of Medicago (Leguminosae). Theor. Appl. Genet. 94, 682–691.

Jeridi, M., Bakry, F., Escoute, J., Fondi, E., Carreel, F., Ferchichi, A., D'Hont, A., and Rodier-Goud, M. (2011). Homoeologous chromosome pairing between the A and B genomes of Musa spp. revealed by genomic in situ hybridization. Ann. Bot.

Jiao, Y., and Paterson, A.H. (2014). Polyploidy-associated genome modifications during land plant evolution. Philos. Trans. R. Soc. B Biol. Sci. 369.

Jin, BGI team, and SEG team (in prep.). The genome sequence of Musa balbisiana.

Jourda, C., Cardi, C., Mbéguié-A-Mbéguié, D., Bocs, S., Garsmeur, O., D'Hont, A., and Yahiaoui, N. (2014). Expansion of banana (Musa acuminata) gene families involved in ethylene biosynthesis and signalling after lineage-specific whole-genome duplications. New Phytol. 202, 986–1000.

Karakoc, E., Alkan, C., O'Roak, B.J., Dennis, M.Y., Vives, L., Mark, K., Rieder, M.J., Nickerson, D.A., and Eichler, E.E. (2012). Detection of structural variants and indels within exome data. Nat. Methods 9, 176–178.

Keane, T.M., Wong, K., and Adams, D.J. (2013). RetroSeq: transposable element discovery from next-generation sequencing data. Bioinformatics 29, 389–390.

Keane, T.M., Wong, K., Adams, D.J., Flint, J., Reymond, A., and Yalcin, B. (2014). Identification of structural variation in mouse genomes. Front. Genet. 5.

Kejnovsky, E., Hawkins, J., and Feschotte, C. (2012). Plant Transposable Elements: Biology and Evolution. In Plant Genome Diversity Volume 1, J.F. Wendel, J. Greilhuber, J. Dolezel, and I.J. Leitch, eds. (Springer Vienna), pp. 17–34.

Kennedy, J. (2009). Bananas and people in the homeland of genus Musa: not just pretty fruit. Ethnobot. Res. Appl. 7, 179–197.

Kim, P.M., Lam, H.Y.K., Urban, A.E., Korbel, J.O., Affourtit, J., Grubert, F., Chen, X., Weissman, S., Snyder, M., and Gerstein, M.B. (2008). Analysis of copy number variants and segmental duplications in the human genome: Evidence for a change in the process of formation in recent evolutionary history. Genome Res. 18, 1865–1874.

Klambauer, G., Schwarzbauer, K., Mayr, A., Clevert, D.-A., Mitterecker, A., Bodenhofer, U., and Hochreiter, S. (2012). cn.MOPS: mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. Nucleic Acids Res. 40, e69–e69.

Knoop, V. (2004). The mitochondrial DNA of land plants: peculiarities in phylogenetic perspective. Curr. Genet. 46, 123–139.

Kohany, O., Gentles, A., Hankus, L., and Jurka, J. (2006). Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. BMC Bioinformatics 7, 474.

Korbel, J.O., Urban, A.E., Affourtit, J.P., Godwin, B., Grubert, F., Simons, J.F., Kim, P.M., Palejev, D., Carriero, N.J., Du, L., et al. (2007). Paired-End Mapping Reveals Extensive Structural Variation in the Human Genome. Science 318, 420–426.

Korbel, J.O., Abyzov, A., Mu, X.J., Carriero, N., Cayting, P., Zhang, Z., Snyder, M., and Gerstein, M.B. (2009). PEMer: a computational framework with simulation-based error

models for inferring genomic structural variants from massive paired-end sequencing data. Genome Biol. 10, R23.

Krumsiek, J., Arnold, R., and Rattei, T. (2007). Gepard: a rapid and sensitive tool for creating dotplots on genome scale. Bioinformatics 23, 1026–1028.

De Langhe, E., Hřibová, E., Carpentier, S., Doležel, J., and Swennen, R. (2010). Did backcrossing contribute to the origin of hybrid edible bananas? Ann. Bot. 106, 849–857.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–359.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 10, R25.

De Lapeyre de Bellaire, L., Fouré, E., Abadie, C., and Carlier, J. (2010). Black Leaf Streak Disease is challenging the banana industry. Fruits 65, 327–342.

Lashermes, P., Combes, M.C., Prakash, N.S., Trouslot, P., Lorieux, M., and Charrier, A. (2001). Genetic linkage map of Coffea canephora: effect of segregation distortion and analysis of recombination rate in male and female meioses. Genome 44, 589–595.

Lassoudière, A. (2007). Bananier et sa culture (le) (Éditions Quae).

Lee, S., Hormozdiari, F., Alkan, C., and Brudno, M. (2009). MoDIL: detecting small indels from clone-end sequencing with mixtures of distributions. Nat. Methods 6, 473–474.

Lescot, T. (2014). La diversité génétique des bananiers. FruiTrop 221, 98.

Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows–Wheeler transform. Bioinformatics 26, 589–595.

Li, L.-F., Häkkinen, M., Yuan, Y.-M., Hao, G., and Ge, X.-J. (2010). Molecular phylogeny and systematics of the banana family (Musaceae) inferred from multiple nuclear and chloroplast DNA fragments, with a special reference to the genus Musa. Mol. Phylogenet. Evol. 57, 1–10.

Li, W., Lin, Z., and Zhang, X. (2007). A Novel Segregation Distortion in Intraspecific Population of Asian Cotton (Gossypium arboretum L.) Detected by Molecular Markers. J. Genet. Genomics 34, 634–640.

Luo, M.C., Deal, K.R., Akhunov, E.D., Akhunova, A.R., Anderson, O.D., Anderson, J.A., Blake, N., Clegg, M.T., Coleman-Derr, D., Conley, E.J., et al. (2009). Genome comparisons reveal a dominant mechanism of chromosome number reduction in grasses and accelerated genome evolution in Triticeae. Proc. Natl. Acad. Sci. 106, 15780–15785.

Lysak, M.A., Berr, A., Pecinka, A., Schmidt, R., McBreen, K., and Schubert, I. (2006). Mechanisms of chromosome number reduction in Arabidopsis thaliana and related Brassicaceae species. Proc. Natl. Acad. Sci. U. S. A. 103, 5224–5229.

Ma, X.-F., Jensen, E., Alexandrov, N., Troukhan, M., Zhang, L., Thomas-Jones, S., Farrar, K., Clifton-Brown, J., Donnison, I., Swaller, T., et al. (2012). High Resolution Genetic Mapping by Genome Sequencing Reveals Genome Duplication and Tetraploid Genetic Structure of the Diploid Miscanthus sinensis. PLoS ONE 7, e33821.

Magi, A., Benelli, M., Yoon, S., Roviello, F., and Torricelli, F. (2011). Detecting common copy number variants in high-throughput sequencing data by using JointSLM algorithm. Nucleic Acids Res. 39, e65–e65.

Mahama, A., Deaderick, L., Sadanaga, K., Newhouse, K., and Palmer, R. (1999). Cytogenetic analysis of translocations in Soybean. J. Hered. 90, 648–653.

Mangelsdorf, P.C., and Jones, D.F. (1926). THE EXPRESSION OF MENDELIAN FACTORS IN THE GAMETOPHYTE OF MAIZE. Genetics 11, 423–455.

Martin, G.E., Rousseau-Gueutin, M., Cordonnier, S., Lima, O., Michon-Coudouel, S., Naquin, D., de Carvalho, J.F., Aïnouche, M., Salmon, A., and Aïnouche, A. (2014). The first complete chloroplast genome of the Genistoid legume Lupinus luteus: evidence for a novel major lineage-specific rearrangement and new insights regarding plastome evolution in the legume family. Ann. Bot.

Mbanjo, E., Tchoumbougnang, F., Mouelle, A., Oben, J., Nyine, M., Dochez, C., Ferguson, M., and Lorenzen, J. (2012). Molecular marker-based genetic linkage map of a diploid banana population (Musa acuminata Colla). Euphytica 188, 369–386.

McClintock, B. (1950). The origin and behavior of mutable loci in maize. Proc. Natl. Acad. Sci. 36, 344–355.

McCouch, S.R., Kochert, G., Yu, Z.H., Wang, Z.Y., Khush, G.S., Coffman, W.R., and Tanksley, S.D. (1988). Molecular mapping of rice chromosomes. Theor. Appl. Genet. 76, 815–829.

McKernan, K.J., Peckham, H.E., Costa, G.L., McLaughlin, S.F., Fu, Y., Tsung, E.F., Clouser, C.R., Duncan, C., Ichikawa, J.K., Lee, C.C., et al. (2009). Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. Genome Res. 19, 1527–1541.

Medvedev, P., Fiume, M., Dzamba, M., Smith, T., and Brudno, M. (2010). Detecting copy number variation with mated short reads. Genome Res. 20, 1613–1622.

Millen, R.S., Olmstead, R.G., Adams, K.L., Palmer, J.D., Lao, N.T., Heggie, L., Kavanagh, T.A., Hibberd, J.M., Gray, J.C., Morden, C.W., et al. (2001). Many Parallel Losses of infA from Chloroplast DNA during Angiosperm Evolution with Multiple Independent Transfers to the Nucleus. Plant Cell Online 13, 645–658.

Mizuta, Y., Harushima, Y., and Kurata, N. (2010). Rice pollen hybrid incompatibility caused by reciprocal gene loss of duplicated genes. Proc. Natl. Acad. Sci. 107, 20417–20422.

Muller, H.J. (1942). Isolating mechanisms, evolution and temperature. In Biol. Symp, pp. 71–125.

Negash, A. (2001). Diversity and conservation of enset (Ensete ventricosum Welw. Cheesman) and its relation to household food and livelihood security in south-western Ethiopia. s.n.].

Newman, T.L., Tuzun, E., Morrison, V.A., Hayden, K.E., Ventura, M., McGrath, S.D., Rocchi, M., and Eichler, E.E. (2005). A genome-wide survey of structural variation between human and chimpanzee. Genome Res. 15, 1344–1356.

Northcott, P.A., Shih, D.J.H., Peacock, J., Garzia, L., Sorana Morrissy, A., Zichner, T., Stutz, A.M., Korshunov, A., Reimand, J., Schumacher, S.E., et al. (2012). Subgroup-specific structural variation across 1,000 medulloblastoma genomes. Nature 488, 49–56.

Notsu, Y., Masood, S., Nishikawa, T., Kubo, N., Akiduki, G., Nakazono, M., Hirai, A., and Kadowaki, K. (2002). The complete sequence of the rice (Oryza sativa L.) mitochondrial genome: frequent DNA sequence acquisition and loss during the evolution of flowering plants. Mol. Genet. Genomics 268, 434–445.

Noumbissié, G.B., Chabannes, M., Bakry, F., Cardi, C., Njembele, J.-C., Yohoume, D., Ricci, S., Tomekpe, K., Iskra-Caruana, M.-L., D'Hont, A., et al. (submitted). Chromosome segregation and eBSV elimination in offsprings from a Musa interspecific tetraploid AAAB.

Novák, P., Hřibová, E., Neumann, P., Koblížková, A., Doležel, J., and Macas, J. (2014). Genome-Wide Analysis of Repeat Diversity across the Family Musaceae. PLoS ONE 9, e98918.

Noyer, J.L., Dambier, D., Lanaud, C., and Lagoda, P. (1997). The saturated map of diploid banana (Musa acuminata). In Abstract Plant & Animal Genome V Conference,.

Noyer, J.L., Causse, S., Tomekpe, K., Bouet, A., and Baurens, F.C. (2005). A new image of plantain diversity assessed by SSR, AFLP and MSAP markers. Genetica 124, 61–69.

Palmer, J. (1992). Mitochondrial DNA in Plant Systematics: Applications and Limitations. In Molecular Systematics of Plants, P. Soltis, D. Soltis, and J. Doyle, eds. (Springer US), pp. 36–49.

Palmer, J.D. (1991). Plastid chromosomes: structure and evolution. In Cell Culture and Somatic Cell Genetics of Plants, L. Bogorad, and I. Vasil, eds. (San Diego: Academic Press), pp. 5–53.

Palmer, J.D., and Thompson, W.F. (1982). Chloroplast DNA rearrangements are more frequent when a large inverted repeat sequence is lost. Cell 29, 537–550.

Palmer, J., and Herbon, L. (1988). Plant mitochondrial DNA evolved rapidly in structure, but slowly in sequence. J. Mol. Evol. 28, 87–97.

Perrier, X., Bakry, F., Carreel, F., Jenny, C., Horry, J.-P., Lebot, V., and Hippolyte, I. (2009). Combining Biological Approaches to Shed Light on the Evolution of Edible Bananas. Ethnobot. Res. Appl. 7.

Perrier, X., De Langhe, E., Donohue, M., Lentfer, C., Vrydaghs, L., Bakry, F., Carreel, F., Hippolyte, I., Horry, J.-P., Jenny, C., et al. (2011). Multidisciplinary perspectives on banana (Musa spp.) domestication. Proc. Natl. Acad. Sci. 108, 11311–11318.

Philippe, R., Choulet, F., Paux, E., van Oeveren, J., Tang, J., Wittenberg, A., Janssen, A., van Eijk, M., Stormo, K., Alberti, A., et al. (2012). Whole Genome Profiling provides a robust framework for physical mapping and sequencing in the highly complex and repetitive wheat genome. BMC Genomics 13, 47.

Ploetz, R.C. (1994). Panama disease: Return of the first banana menace. Int. J. Pest Manag. 40, 326–336.

Pop, M. (2009). Genome assembly reborn: recent computational challenges. Brief. Bioinform. 10, 354–366.

Pungetti, G., and MacIvor, A. (2007). Preliminary Literature Review on Sacred Species.

Qi, J., and Zhao, F. (2011). inGAP-sv: a novel scheme to identify and visualize structural variation from paired end mapping data. Nucleic Acids Res. 39, W567–W575.

Quillet, M.C., Madjidian, N., Griveau, Y., Serieys, H., Tersac, M., Lorieux, M., and Bervillé, A. (1995). Mapping genetic factors controlling pollen viability in an interspecific cross in Helianthus sect. Helianthus. Theor. Appl. Genet. 91, 1195–1202.

Quinlan, A.R., Clark, R.A., Sokolova, S., Leibowitz, M.L., Zhang, Y., Hurles, M.E., Mell, J.C., and Hall, I.M. (2010). Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome. Genome Res. 20, 623–635.

Raboin, L.M., Carreel, F., Noyer, J.-L., Baurens, F.-C., Horry, J.P., Bakry, F., Tézenas du Montcel, H., Ganry, J., Lanaud, C., and Lagoda, P.J.L. (2005). Diploid ancestors of triploid export banana cultivars: molecular identification of 2n restitution gamete donors and n gamete donors. Mol. Breed. 16, 333–341.

Ramsey, J., and Schemske, D.W. (1998). Pathways, Mechanisms, And Rates Of Polyploid Formation In Flowering Plants. Annu. Rev. Ecol. Syst. 29, 467–501.

Raubeson, L.A., and Jansen, R.K. (2005). Chloroplast genomes of plants. In Plant Diversity and Evolution: Genotypic and Phenotypic Variation in Higher Plants, R.J. Henry, ed. (Cambridge: CAB International), pp. 45–68.

Rausch, T., Zichner, T., Schlattl, A., Stutz, A.M., Benes, V., and Korbel, J.O. (2012). DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics 28, i333–i339.

Rens, W., Grützner, F., O'Brien, P.C.M., Fairclough, H., Graves, J.A.M., and Ferguson-Smith, M.A. (2004). Resolution and evolution of the duck-billed platypus karyotype with an X1Y1X2Y2X3Y3X4Y4X5Y5 male sex chromosome constitution. Proc. Natl. Acad. Sci. U. S. A. 101, 16257–16261.

Rieseberg, L.H. (2001). Chromosomal rearrangements and speciation. Trends Ecol. Evol. 16, 351–358.

Risterucci, A.-M., Hippolyte, I., Perrier, X., Xia, L., Caig, V., Evers, M., Huttner, E., Kilian, A., and Glaszmann, J.-C. (2009). Development and assessment of Diversity Arrays Technology for high-throughput DNA analyses in Musa. Theor. Appl. Genet. 119, 1093–1103.

Rizk, G., Gouin, A., Chikhi, R., and Lemaitre, C. (2014). MindTheGap : integrated detection and assembly of short and long insertions. Bioinformatics.

Sabot, F., Picault, N., El-Baidouri, M., Llauro, C., Chaparro, C., Piegu, B., Roulin, A., Guiderdoni, E., Delabastide, M., McCombie, R., et al. (2011). Transpositional landscape of the rice genome revealed by paired-end mapping of high-throughput re-sequencing data. Plant J. 66, 241–246.

Schatz, M., Witkowski, J., and McCombie, W.R. (2012). Current challenges in de novo plant genome sequencing and assembly. Genome Biol. 13, 243.

Schnable, J.C., Springer, N.M., and Freeling, M. (2011). Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. Proc. Natl. Acad. Sci. 108, 4069–4074.

Sequencing Project International Rice Genome (2005). The map-based sequence of the rice genome. Nature 436, 793–800.

Shepherd, K. (1999). Cytogenetics of the genus Musa (IPGRI).

Shetty, S., Griffin, D., and Graves, J.M. (1999). Comparative Painting Reveals Strong Chromosome Homology Over 80 Million Years of Bird Evolution. Chromosome Res. 7, 289–295.

Simmonds, N.W. (1959). Bananas.

Simmonds, N.W. (1962). The evolution of the bananas (London: Longmans).

Simmonds, N.W., and Shepherd, K. (1955). The taxonomy and origins of the cultivated bananas. J. Linn. Soc. Lond. Bot. 55, 302–312.

Simpson, J.T., McIntyre, R.E., Adams, D.J., and Durbin, R. (2010). Copy number variant detection in inbred strains from short read sequence data. Bioinformatics 26, 565–567.

Sindi, S.S., Önal, S., Peng, L.C., Wu, H.-T., and Raphael, B.J. (2012). An integrative probabilistic model for identification of structural variation in sequencing data. Genome Biol. 13, R22.

Soltis, P.S., Liu, X., Marchant, D.B., Visger, C.J., and Soltis, D.E. (2014). Polyploidy and novelty: Gottlieb's legacy. Philos. Trans. R. Soc. B Biol. Sci. 369.

Spindel, J., Wright, M., Chen, C., Cobb, J., Gage, J., Harrington, S., Lorieux, M., Ahmadi, N., and McCouch, S. (2013). Bridging the genotyping gap: using genotyping by sequencing (GBS) to add high-density SNP markers and new value to traditional bi-parental mapping and breeding populations. Theor. Appl. Genet. 1–18.

Sweigart, A.L., and Willis, J.H. (2012). Molecular evolution and genetics of postzygotic reproductive isolation in plants. F1000 Biol. Rep. 4.

Tadmor, Y., Zamir, D., and Ladizinsky, G. (1987). Genetic mapping of an ancient translocation in the genus Lens. Theor. Appl. Genet. 73, 883–892.

Takumi, S., Motomura, Y., Iehisa, J., and Kobayashi, F. (2013). Segregation distortion caused by weak hybrid necrosis in recombinant inbred lines of common wheat. Genetica 141, 463–470.

Talukdar, D. (2010). Reciprocal Translocations in Grass Pea (Lathyrus sativus L.): Pattern of Transmission, Detection of Multiple Interchanges and their Independence. J. Hered. 101, 169–176.

The 1000 Genomes Project Consortium (2010). A map of human genome variation from population-scale sequencing. Nature 467, 1061–1073.

The 1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. Nature 491, 56–65.

The Arabidopsis Genome Initiative (2000). Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature 408, 796–815.

Thomas, B.C., Pedersen, B., and Freeling, M. (2006). Following tetraploidy in an Arabidopsis ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes. Genome Res. 16, 934–946.

Toder, R., O'Neill, R.W., Wienberg, J., O'Brien, P.M., Voullaire, L., and Marshall-Graves, J. (1997). Comparative chromosome painting between two marsupials: origins of an XX/XY1Y2 sex chromosome system. Mamm. Genome 8, 418–422.

Torres, A.M., Mau-Lastovicka, T., Williams, T.E., and Soost, R.K. (1985). Segregation distortion and linkage of Citrus and Poncirus isozyme genes. J. Hered. 76, 289–294.

Tuzun, E., Sharp, A.J., Bailey, J.A., Kaul, R., Morrison, V.A., Pertz, L.M., Haugen, E., Hayden, H., Albertson, D., Pinkel, D., et al. (2005). Fine-scale structural variation of the human genome. Nat. Genet. 37, 727–732.

Vanneste, K., Maere, S., and Van de Peer, Y. (2014). Tangled up in two: a burst of genome duplications at the end of the Cretaceous and the consequences for plant evolution. Philos. Trans. R. Soc. B Biol. Sci. 369.

Vilarinhos, A.D. (2004). Cartographie génétique et cytogénétique chez le bananier: caractérisation des translocations.

Wang, J., Mullighan, C.G., Easton, J., Roberts, S., Heatley, S.L., Ma, J., Rusch, M.C., Chen, K., Harris, C.C., Ding, L., et al. (2011). CREST maps somatic structural variation in cancer genomes with base-pair resolution. Nat. Methods 8, 652–654.

White, F. (1962). Geographic variation and speciation in Africa with particular reference to Diospyros. In Taxonomy and Geography, (London: Nichols D), pp. 71–103.

Wong, C., Kiew, R., Argent, G., Set, O., Lee, S.K., and Gan, Y.Y. (2002). Assessment of the Validity of the Sections in Musa (Musaceae) using AFLP. Ann. Bot. 90, 231–238.

Wong, K., Keane, T.M., Stalker, J., and Adams, D.J. (2010). Enhanced structural variant and breakpoint detection using SVMerge by integration of multiple detection methods and local assembly. Genome Biol. 11, R128.

Woodhouse, M.R., Schnable, J.C., Pedersen, B.S., Lyons, E., Lisch, D., Subramaniam, S., and Freeling, M. (2010). Following Tetraploidy in Maize, a Short Deletion Mechanism Removed Genes Preferentially from One of the Two Homeologs. PLoS Biol. 8, e1000409.

Wu, J., Mizuno, H., Hayashi-Tsugane, M., Ito, Y., Chiden, Y., Fujisawa, M., Katagiri, S., Saji, S., Yoshiki, S., Karasawa, W., et al. (2003). Physical maps and recombination frequency of six rice chromosomes. Plant J. 36, 720–730.

Xi, R., Kim, T.-M., and Park, P.J. (2010). Detecting structural variations in the human genome using next generation sequencing. Brief. Funct. Genomics 9, 405–415.

Xie, C., and Tammi, M.T. (2009). CNV-seq, a new method to detect copy number variation using high-throughput sequencing. BMC Bioinformatics 10, 80.

Xie, S., Ramanna, M., Visser, R.F., Arens, P., and van Tuyl, J. (2013). Elucidation of intergenomic recombination and chromosome translocation: meiotic evidence from interspecific hybrids of Lilium through GISH analysis. Euphytica 194, 361–370.

Xing, J., Zhang, Y., Han, K., Salem, A.H., Sen, S.K., Huff, C.D., Zhou, Q., Kirkness, E.F., Levy, S., Batzer, M.A., et al. (2009). Mobile elements create structural variation: Analysis of a complete human genome. Genome Res. 19, 1516–1526.

Yang, M., Zhang, X., Liu, G., Yin, Y., Chen, K., Yun, Q., Zhao, D., Al-Mssallem, I.S., and Yu, J. (2010). The Complete Chloroplast Genome Sequence of Date Palm (Phoenix dactylifera L.). PLoS ONE 5, e12762.

Ye, K., Schulz, M.H., Long, Q., Apweiler, R., and Ning, Z. (2009). Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics 25, 2865–2871.

Yoon, S., Xuan, Z., Makarov, V., Ye, K., and Sebat, J. (2009). Sensitive and accurate detection of copy number variants using read depth of coverage. Genome Res. 19, 1586–1592.

Zeitouni, B., Boeva, V., Janoueix-Lerosey, I., Loeillet, S., Legoix-né, P., Nicolas, A., Delattre, O., and Barillot, E. (2010). SVDetect: a tool to identify genomic structural variations from paired-end and mate-pair sequencing data. Bioinformatics 26, 1895–1896.

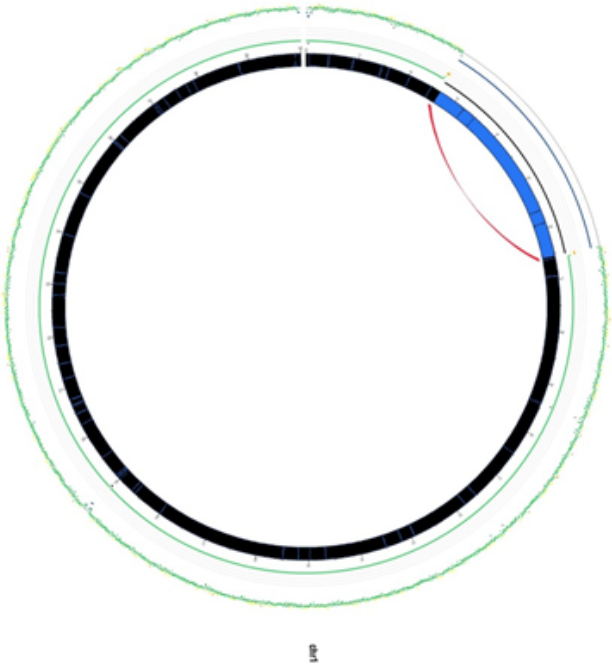
Zhang, J., Wang, J., and Wu, Y. (2012). An improved approach for accurate and efficient calling of structural variations with low-coverage sequence data. BMC Bioinformatics 13, S6.

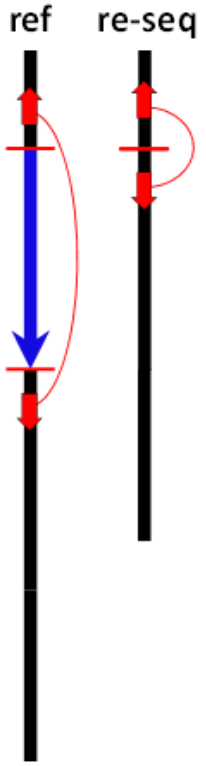
Zhang, Z.D., Du, J., Lam, H., Abyzov, A., Urban, A.E., Snyder, M., and Gerstein, M. (2011). Identification of genomic indels and structural variations using split reads. BMC Genomics 12, 375.

Zhu, X.-Y., Chase, M., Qiu, Y.-L., Kong, H.-Z., Dilcher, D., Li, J.-H., and Chen, Z.-D. (2007). Mitochondrial matR sequences help to resolve deep phylogenetic relationships in rosids. BMC Evol. Biol. 7, 217.

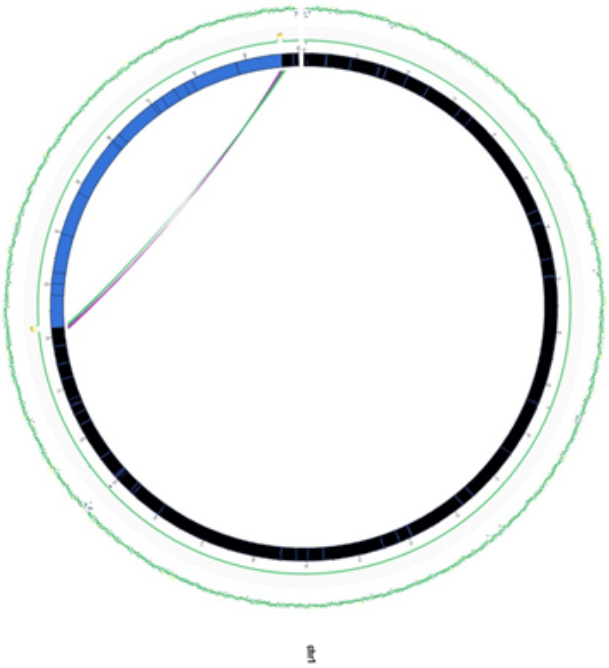
Zivy, M., Devaux, P., Blaisonneau, J., Jean, R., and Thiellement, H. (1992). Segregation distortion and linkage studies in microspore-derived double haploid lines of Hordeum vulgare L. Theor. Appl. Genet. 83, 919–924.

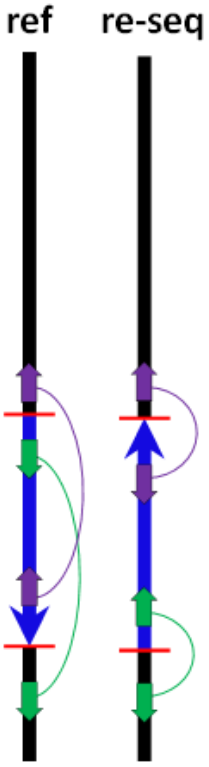
a - Deletion





b - Inversion

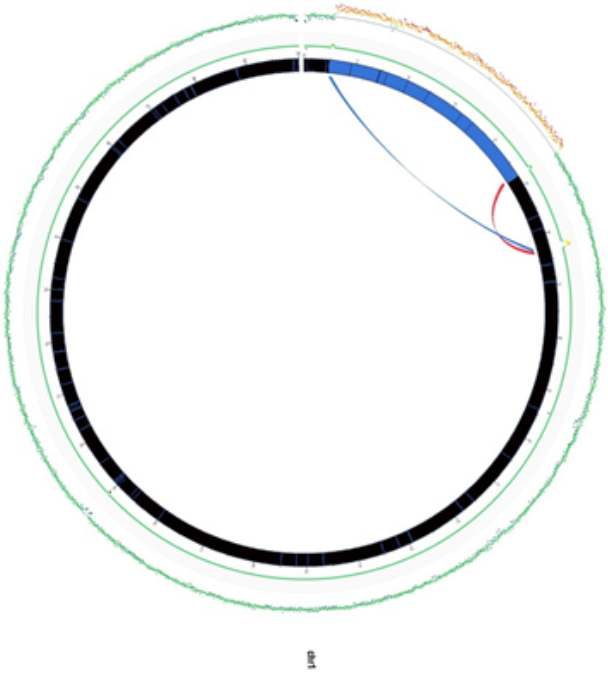


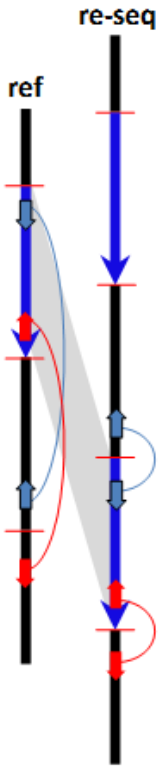


Annexes

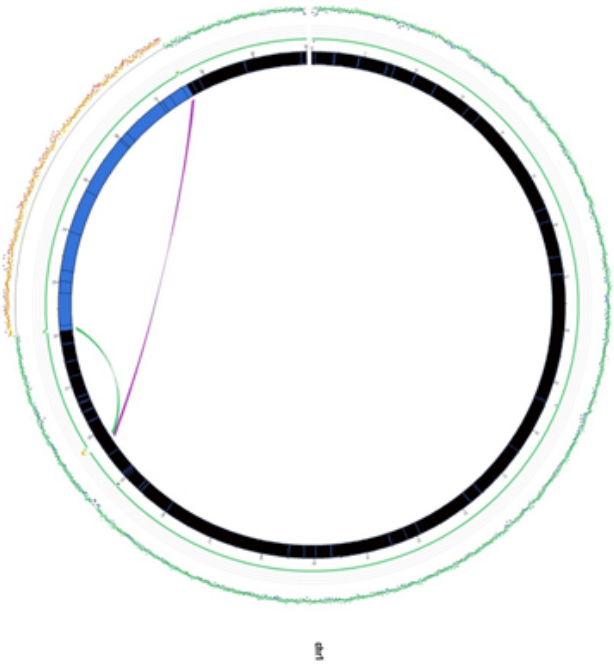
Annexe 1 : Signatures de variations structurales recherchées par le pipeline présenté dans le chapitre I. Le pipeline recherche différents types de variations structurales: (a) délétion, (b) inversion, (c) duplication, (d) duplication avec inversion du fragment dupliqué, (e) translocation, (f) translocation avec inversion du fragment transloqué, (g) translocation réciproque, (h) translocation réciproque avec inversion des deux fragments transloqués, (i) translocation réciproque avec inversion d'un des fragments transloqué et (j) translocation réciproque avec inversion de l'autre fragment transloqué par rapport au (i). Pour chacune des variations structurales recherchées un schéma explique comment s'alignent sur le génome de référence (ref) les lectures pairées obtenues (flèches) sur une accession re-séquencée (re-seq) qui présente la variation. Pour chacune des variations structurales recherchées une représentation Circos des paires de zones de discordances attendues est réalisée. Le code couleur entre le schéma et la figure circos est les même et suit le code couleur du pipeline.

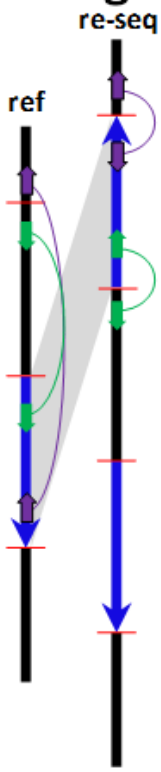
c - Duplication



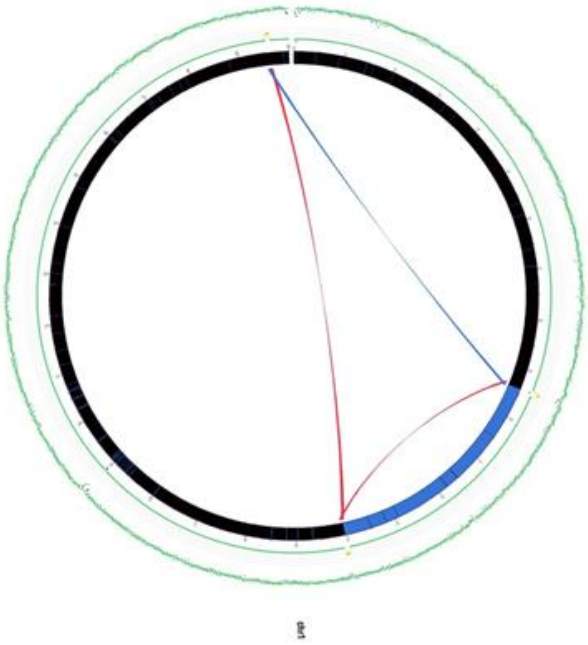


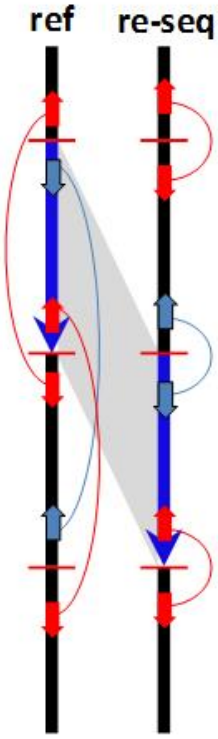
d - Duplication with inverted duplicated fragment



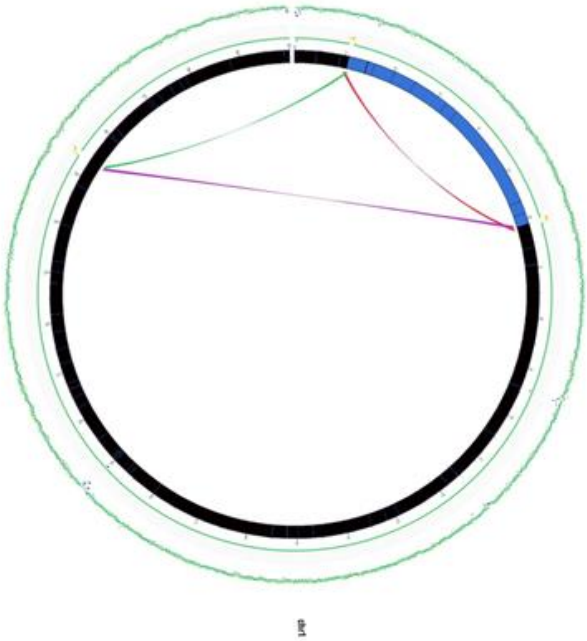


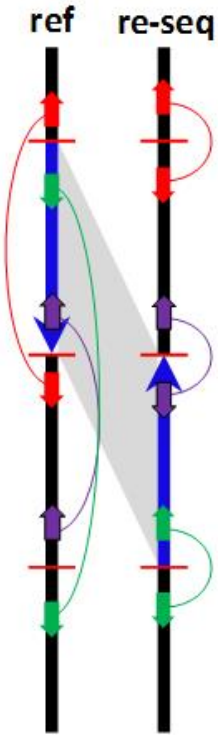
e - Translocation



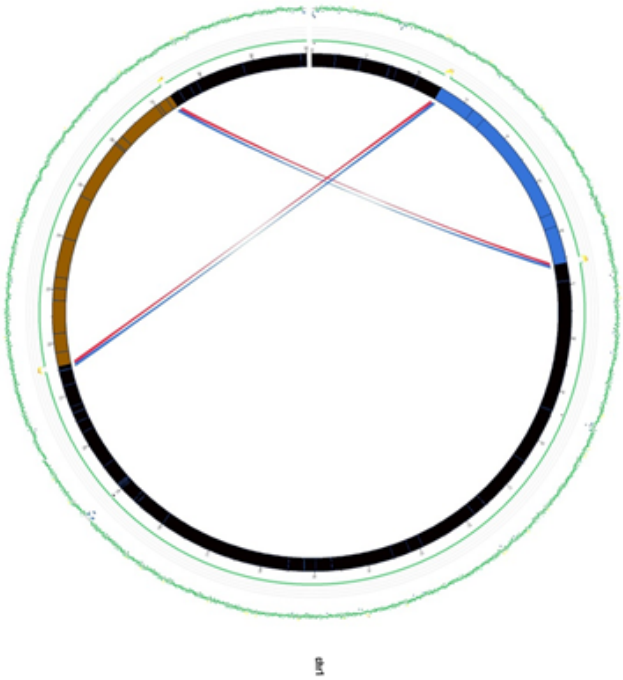


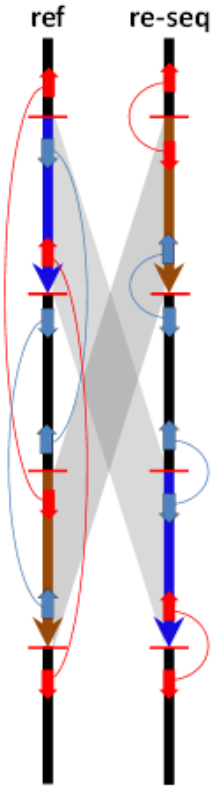
f - Translocation with inverted duplicated fragment



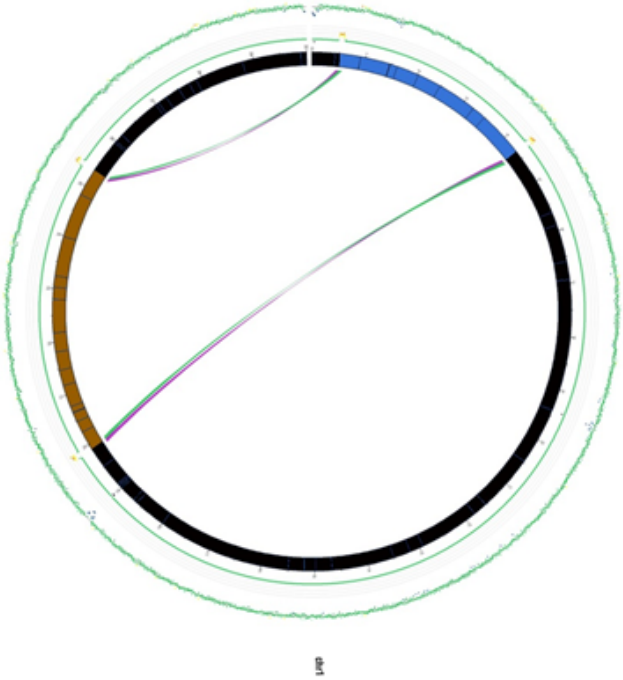


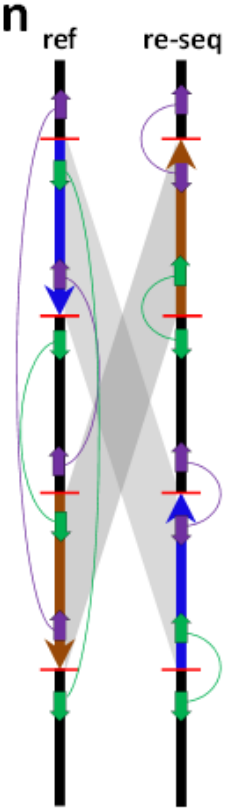
g - Reciprocal translocation and no inversion of translocated fragments



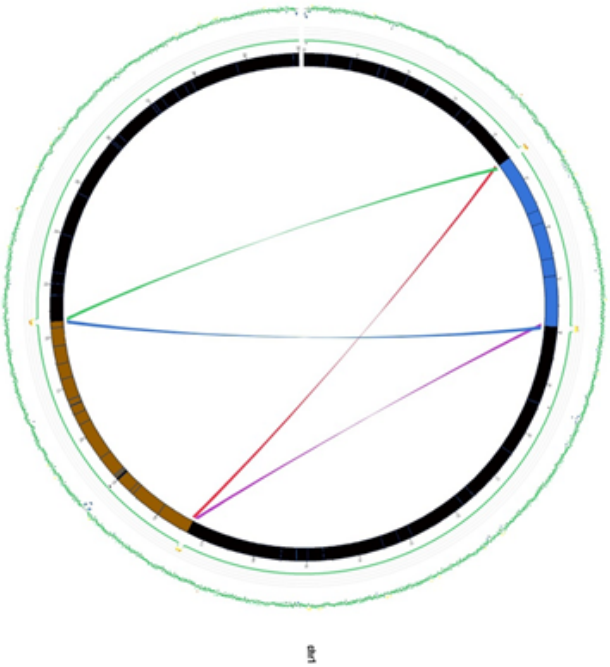


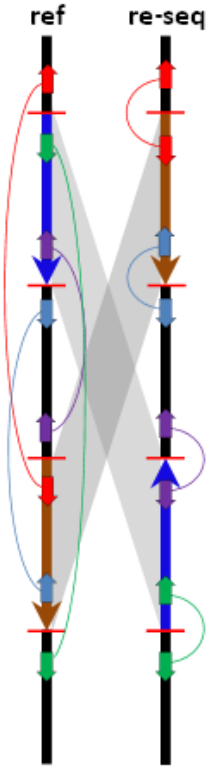
h - Reciprocal translocation and inversion of both translocated fragments



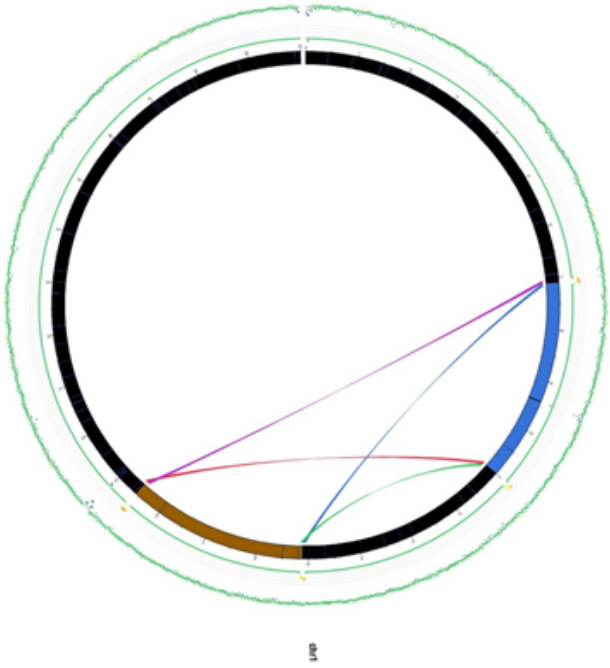


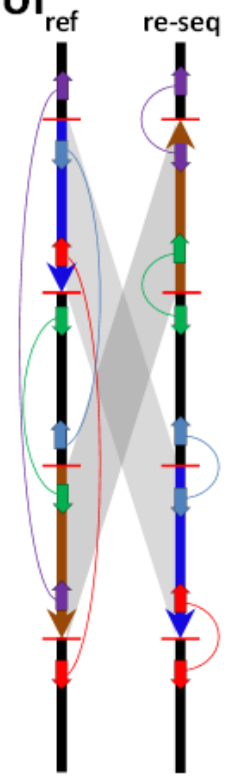
i - Reciprocal translocation and inversion of first translocated fragment





j - Reciprocal translocation and inversion of second translocated fragments





Résumé
Les cultivars de bananiers sont dérivés d'hybridations entre sous-espèces de *Musa acuminata* (génome A) et pour certains avec l'espèce *M. balbisiana* (génome B). Ces hybrides présentent une fertilité réduite, des méioses perturbées et de fortes distorsions de ségrégation. Ces caractéristiques attribuées à des réarrangements chromosomiques entre espèces et sous-espèces compliquent les analyses génétiques et les programmes d'amélioration variétale. Au cours de cette thèse, nous avons mis en place et testé de nouvelles approches, basées sur la récente disponibilité d'une séquence de référence du bananier et des technologies de séquençage haut-débit, pour caractériser ces différences de structures chromosomiques et comprendre leur impact sur les ségrégations chromosomiques. Ces approches ont nécessité l'amélioration de la séquence de référence du bananier. Pour cela, des outils ont été développés. Ils sont applicables à d'autres génomes et modulables en fonction des données disponibles. Le nombre de scaffolds a été divisé par 5 et 90% de la séquence est maintenant ancré aux chromosomes. Les scaffolds correspondant au génome mitochondrial ont été identifiés et le génome chloroplastique a été assemblé et annoté. Des données de re-séquençage de l'accession 'Pahang' et de génotypage dense de sa descendance ont été utilisées pour explorer l'origine des distorsions de ségrégation impliquant les chromosomes 1 et 4. L'ensemble des données (profils de distorsion et de recombinaison, appariements à la méiose, re-séquençage), nous orientent vers l'hypothèse d'une translocation réciproque en orientation inversée, entre régions distales des chromosomes 1 et 4. Le test de nos outils de recherche de variations structurales pour comparer les génomes A et B du bananier, dont les différences de structure sont connues, montre que nos outils détectent directement les signatures de certaines variations structurales mais que pour d'autres il ne détecte que des signatures partielles. Ces dernières peuvent néanmoins être informatives en complément d'autres types d'informations provenant de cartographie génétique et d'analyses cytogénétiques.

Mots clé : *Musa*, variations structurales, distorsions de ségrégation, bioinformatique, re-séquençage, génotypage par séquençage.

Summary
Banana cultivars are derived from hybridization between *Musa acuminata* subspecies (A genome) and, for some of them, with the species *M. balbisiana* (B genome). These hybrids have reduced fertility, disturbed meiosis and strong segregation distortions. These characteristics attributed to chromosomal rearrangements between species and subspecies complicate genetic analyses and breeding programs. In this thesis, we have developed and tested new approaches based on the recent availability of a banana reference genome sequence and high-throughput sequencing technologies, to characterize these differences in chromosomal structures and understand their impact on chromosomal segregation. These approaches needed improvement of the banana reference genome sequence. New bioinformatics tools were developed for this purpose. They are applicable to other genomes and are flexible according to available data. The scaffolds number was divided by 5 and 90% of the assembly is now anchored to the chromosomes. Scaffolds corresponding to the mitochondrial genome were identified and the chloroplast genome was assembled and annotated. Re-sequencing data from the 'Pahang' accession and dense genotyping of its progeny were used to explore the origin of segregation distortion involving chromosomes 1 and 4. Distortion and recombination profiles, chromosomal pairing at meiosis and re-sequencing data direct us to the hypothesis of a reciprocal translocation in inverted orientation between distal portions of chromosomes 1 and 4. We tested our structural variation research tools to compare the A and B genomes of banana, for which structural differences are known. The results showed that our tools detected complete signatures of some structural changes but for others, they only detected partial signatures. The latter can still be informative in addition to other informations derived from genetic mapping and cytogenetic studies.

Key words: *Musa*, structural variations, segregation distortions, bioinformatics, re-sequencing, genotyping by sequencing.